Stabilität bei numerischer Integration von Differentialgleichungen

Inaugural-Dissertation
zur Erlangung des Grades eines Doktors
der Naturwissenschaften
der
Mathematisch-naturwissenschaftlichen Fakultät
der
Eberhard-Karls-Universität
zu Tübingen

vorgelegt von
Manfred Reimer
aus Breslau

1963

7433

Stabilität bei numerischer Integration von Differentialgleichungen

Inaugural-Dissertation
zur Erlangung des Grades eines Doktors
der Naturwissenschaften
der
Mathematisch-naturwissenschaftlichen Fakultät
der
Eberhard-Karls-Universität
zu Tübingen

vorgelegt von
Manfred Reimer
aus Breslau

1963

Tag der mündlichen Prüfung: 14. Februar 1963

Dekan:	Professor Dr. K. G. Grell
1. Berichterstatter	Professor Dr. K. Zeller
2. Berichterstatter	Professor Dr. H. Kneser

Herrn Professor Dr. Karl Zeller
danke ich für die Anregung zu
dieser Arbeit

LEIDEN

Inhaltsverzeichnis.

Kapitel I : Einführung

§ 1 Grundbegriffe ..... Seite 2

§ 2 Stabilität 3

§ 3 Optimale Verfahren, Aufgabenstellung 6

Kapitel II : Zur Darstellung linearer Differenzenverfahren L

§ 4 Taylorentwicklung und Restglied 7

§ 5 Die Klasse $\mathfrak{L}_{k+1}$ 8

§ 6 Die Klasse $\mathfrak{L}_k$ (offene Operatoren) 11

Kapitel III: Absolute Stabilität

§ 7 Vorbemerkungen 14

§ 8 Die Variationsgleichung 16

§ 9 Definitionen. Kriterien für absolute Stabilität 17

§ 10 Weitere Sätze über absolut stabile L 25

Kapitel IV : Optimale Integrationsverfahren

§ 11 Vorbemerkungen 28

§ 12 Maximale Schrittweite 29

§ 13 Der optimale Operator 2.Ordnung 31

§ 14 Der optimale Operator 3.Ordnung 36

§ 15 Optimale Operatoren für $h \to 0$ 48

Literaturverzeichnis 53

K A P I T E L I

Einführung.

§1. Grundbegriffe.

Betrachtet wird ein Anfangswertproblem der Gestalt

(1.1) $\qquad y' = f(x,y), \quad y(a) = y_0, \quad a \leq x \leq b$

mit $a < b$. Es wird dabei vorausgesetzt, daß die Funktion $f = f(x,y)$ mit einer natürlichen Zahl q zur Klasse $\mathcal{F}_q$ gehört. Das bedeutet, daß f in einem Gebiet G der (x,y)-Ebene erklärt und q mal stetig partiell differenzierbar sein soll und daß (1.1) eine (dann sicher eindeutig bestimmte) Lösung $y = y(x)$ besitzt, für die $(x,y(x)) \in G$ für $x \in [a,b]$ gilt.

Es sei für ein festes $h > 0$

$$x_m = a + mh$$

für $m = 0,1,\ldots$. Ist $(x_{n+\nu}, u_{n+\nu}) \in G$ für $\nu = 0,\ldots,k$, so wird durch

(1.2) $\qquad L u_n = \sum_{\nu=0}^{k} \{ a_\nu u_{n+\nu} - h\, b_\nu f(x_{n+\nu}, u_{n+\nu}) \}$

ein linearer Differenzenoperator L definiert. Dieser bestimmt seinerseits im allgemeinen ein lineares Differenzenverfahren [1]

(1.3) $\qquad u_{n+k} = \Lambda(u_n,\ldots,u_{n+k-1}),$

und zwar als Lösung der Differenzengleichung

(1.4) $\qquad L u_n = 0.$

Die a_ν und b_ν $(\nu = 0,\ldots,k)$ heißen die K o e f f i z i e n t e n des Verfahrens und sollen reell sein. Außerdem wird $a_k \neq 0$ vorausgesetzt. Mit Rücksicht auf (1.4) kann dann sogar o.B.d.A.

(1.5) $\qquad a_k = 1$

angenommen werden.

---

1) Wenn keine Mißverständnisse zu befürchten sind, wird auch das Verfahren mit L bezeichnet werden.

Die in (1.2) vorkommende Größe h sei stets positiv und heiße die Schrittweite des Verfahrens, die natürliche Zahl k seine Ordnung.
Man kann die Operatoren L mit gleichen Koeffizienten, aber verschiedenen Schrittweiten zu einer Klasse zusammenfassen und mit dieser Klasse identifizieren. Dementsprechend soll auch von dem Operator L gesprochen werden, wenn seine Klasse gemeint ist.
Setzt man nun

$$y_m = y(x_m) \qquad (m = 0,1,\ldots),$$

so gilt unter gewissen Voraussetzungen über die Koeffizienten von L mit einer geeigneten natürlichen Zahl p

(1.6) $$t_n = L y_n = O(h^{p+1}) \quad \text{für alle } f \in \mathcal{F}_p \quad (h \to 0).$$

Wir nehmen an, p sei die größte natürliche Zahl dieser Art. Dann heiße p der Grad des Operators, während t_n wegen (1.4) der Formelfehler des Verfahrens genannt werden soll.

Man setze $U_\nu = f(x_\nu, u_\nu)$, $V_\nu = f(x_\nu, v_\nu)$ usw., und es sei

(1.7) $$\varrho(s) = \sum_{\nu=0}^{k} a_\nu s^\nu , \quad \sigma(s) = \sum_{\nu=0}^{k} b_\nu s^\nu .$$

Versteht man dann unter E den durch

$$E u_\nu = u_{\nu+1}$$

gekennzeichneten Operator, so kann (1.4) auch in der Form

(1.8) $$\varrho(E) u_n - h \sigma(E) U_n = 0$$

geschrieben werden.

§2. Stabilität.

Die Lösung der Differenzengleichung (1.4) hat nicht immer etwas mit dem Integral von (1.1) zu tun. Vorläufig soll ein Operator L stabil heißen, wenn er erzwingt, daß die numerische Lösung von (1.4) unter geeigneten Voraussetzungen über die Größe der Schrittweite und die Genauigkeit der numerischen Rechnung das Integral von (1.1) beliebig gut approximiert.
Eine Minimalforderung, die an einen stabilen Operator zu stellen ist, ist die, daß die Integrale der Differentialgleichungen der

Gestalt $y' = A = \text{const}$ der Differenzengleichung (1.4) bzw. (1.8) exakt genügen. Aus dieser Forderung ergibt sich die für Stabilität notwendige Bedingung

(2.1) $\qquad \varrho(1) = 0, \quad \varrho'(1) = \sigma(1).$

Diese stimmt mit der Bedingung, L solle einen Grad $p \geq 1$ haben überein und soll im weiteren stets als erfüllt vorausgesetzt werden.[1] Im übrigen aber gibt es eine Fülle von Möglichkeiten, den Begriff der Stabilität zu präzisieren.

Wie H.Rutishauser 1952 zeigt, ist die Lösung von (1.4) im Falle der Differentialgleichung $y' = gy$ mit $g = \text{const}$ im allgemeinen gegeben durch eine Potenzsumme

(2.2) $\qquad u_n = \sum_{i=1}^{k} A_i \sigma_i^n \qquad (n = 0,1,\ldots),$

bei der die $\sigma_i = \sigma_i(h)$ die Wurzeln der charakteristischen Gleichung

(2.3) $\qquad \varrho(s) - gh\sigma(s) = 0$

sind und die Koeffizienten A_i durch die σ_i und die Anfangswerte $u_0,\ldots,u_{k-1}$ bestimmt sind. Bezeichnet man die Wurzeln von $\varrho(s)$ mit s_i $(i = 1,\ldots,k)$, so kann wegen (2.1) o.B.d.A. $s_1 = 1$ angenommen werden, und man hat

(2.4) $\qquad \sigma_1(h) \to 1 \quad$ für $h \to 0.$

Rutishauser bemerkte, daß nur die Komponente $A_1\sigma_1^n$ von u_n etwas mit dem Integral, d.h. mit y_n, zu tun hat, während die anderen Komponenten durch das Verfahren *eingeschleppt* sind (Parasitlösungen) und, wenn $|\sigma_i| > 1$ ist, zu instabilem Verhalten der numerischen Lösung Anlaß geben ("Aufrauhungserscheinung").
Nach Rutishauser hängt also das Stabilitätsverhalten im wesentlichen von den charakteristischen Wurzeln σ_i ab und nicht - wie man wegen (1.6) meinen könnte - von der Größe von p.
Wegen $\sigma_i(h) \sim s_i$ für $h \to 0$ und des oben Gesagten nennt Rutishauser ein Verfahren *instabil*, wenn $|s_i| > 1$ für ein $i \neq 1$, und *stabil*, wenn $|s_i| < 1$ für $i = 2,\ldots,k$ ist. Der Fall, daß alle s_i auf der abgeschlossenen Einheitskreisscheibe liegen und etwa $|s_2| = 1$ ist, wird nicht erfaßt.

---

1) vgl. S.7 (4.5) und (4.6).

Es sei hier noch ausdrücklich darauf hingewiesen, daß von Rutishauser der Fall betrachtet wird, daß bei - zwar vielleicht kleinem, jedoch - festem $h > 0$ über eine große Anzahl von Schritten integriert wird, ferner, daß nur die Auswirkungen einer einmaligen numerischen Störung in den Anfangswerten erörtert werden und nicht der allgemeine Fall von Störungen bei jedem Schritt. Denn in diesem letzteren Fall würden die Koeffizienten A_i in (2.2) von n abhängen.

G. Dahlquist 1956 nennt einen Operator L mit $(\varrho,\sigma) = 1$ [1] genau dann stabil, wenn für $1 \leq i \leq k$

(D) $\qquad p \geq 1, \quad |s_i| \leq 1, \quad \varrho'(s_i) \neq 0 \quad \text{für} |s_i| = 1$

gilt, und zeigt, daß die Bedingung (D) notwendig und hinreichend ist für "stabile Konvergenz". Stabile Konvergenz bedeutet dabei - grob gesagt - gleichmäßige Konvergenz der numerischen Lösung von (1.4) im Intervall $a \leq x \leq b$ für $h \to 0$ unter der Nebenbedingung $n = \left[\frac{x-a}{h}\right]$, wenn die Anfangswertfehler und die lokalen Störungen (Runde- und Abbrechfehler) von der Größenordnung $O(h^\beta)$ mit einem $\beta > 1$ sind. Wir heben hervor, daß der Stabilitätsbegriff von Dahlquist sich auf das endliche Integrationsintervall bezieht und die in der Praxis vorliegende Situation vollständig erfaßt.

Außer den grundlegenden Arbeiten von H.Rutishauser und G.Dahlquist über Fragen der numerischen Stabilität bei linearen Differenzenverfahren ist noch eine Reihe weiterer erschienen, von denen hier nur einige genannt werden sollen.
W.Quade 1957 nennt ein Verfahren stabil, wenn (in Abhängigkeit von Störungen der Anfangswerte) alle Lösungen der Variationsgleichung zu (1.4) für $n \to \infty$ Nullfolgen sind - ein Stabilitätsbegriff, der weitgehend mit dem Rutishausers übereinstimmt.
Bei W.Liniger 1957 heißt ein Operator bereits dann stabil, wenn es eine Schrittweite h und eine Integralkurve y(x) gibt,

---

1) Diese Bedingung ist beweistechnisch angenehm und stellt nur sicher, daß L nicht durch Operatoren niederer Ordnung ersetzt werden kann.

so daß alle Lösungen der Variationsgleichung zu (1.4), genommen für $y(x)$ und h, beschränkt sind. Dieser Begriff der Stabilität erscheint uns vom Standpunkt der Praxis aus zu allgemein. T.E.Hull und A.C.R.Newbery 1962 nennen ein Verfahren L relativ stabil, wenn $|\sigma_i(h)| < |\sigma_1(h)|$ für $i = 2,\ldots,k$ ausfällt, d.h. wenn die Parasitlösungen für $n \to \infty$ ein schwächeres Wachstum aufweisen als das Integral. Dieser Stabilitätsbegriff ist sicher in Bezug auf instabile Differentialgleichungen ($f_y < 0$) bedeutungsvoll. Man vergleiche hierzu auch die Klassifikation der Verfahren bei F.Krückeberg 1961 S.83.

§3. Optimale Verfahren, Aufgabenstellung.

Im weiteren werden wir nur auf den Stabilitätsbegriff von Dahlquist zurückgreifen und kurz von D-Stabilität sprechen.
Wie man an der Bedingung (D) erkennt, gibt es eine Fülle D-stabiler Operatoren L. Somit stellt sich die Aufgabe, den Stabilitätsbegriff so einzuengen, daß möglichst genau ein Verfahren vorgeschriebener Ordnung in dem neuen Sinne stabil ist. Dieses Verfahren könnte dann als das beste seiner Ordnung bezeichnet werden.
Eine Einengung des Stabilitätsbegriffes von Dahlquist wird von uns durch Einführung des Begriffes der absoluten Stabilität in Kapitel III gegeben. Dieser bestimmt aber noch keinen "besten" Operator in der oben geforderten Weise. In Kapitel IV wird die Frage nach dem besten Operator für stabile Differentialgleichungen behandelt und für die Fälle $k = 2$ und 3 zufriedenstellend beantwortet. Es zeigt sich dabei, daß der optimale Operator nicht unabhängig von der Differentialgleichung bestimmt werden kann, es sei denn, daß "unbegrenzte Genauigkeit" ($h \to 0$) verlangt wird. Zuvor aber werden in Kapitel II zwei wichtige Klassen von linearen Differenzenverfahren L explizit dargestellt durch Formeln, die eine einfache numerische Erzeugung der Operatoren ermöglichen.

K A P I T E L II

Zur Darstellung linearer Differenzenverfahren L.

§4. Taylorentwicklung und Restglied.

Es gelten die Bezeichnungen und Voraussetzungen von Kapitel I. Insbesondere sei stets $p \geq 1$.[1)]

Entwickelt man $L y_n$ an der Stelle x_n in eine Taylorreihe, so erhält man mit $y_n^{(i)} = y^{(i)}(x_n)$

$$L y_n = \sum_{i=0}^{q} \frac{1}{i!} c_i \, y_n^{(i)} h^i + R_{n,q+1} \quad \text{für} \quad f \in \mathfrak{F}_q . \tag{4.1}$$

Die c_i können durch die Koeffizienten des Verfahrens ausgedrückt werden. Man findet

$$c_i = \sum_{\varkappa=0}^{k} (\varkappa^i \cdot a_\varkappa - i \cdot \varkappa^{i-1} \cdot b_\varkappa) \qquad (i = 0,1,\ldots). \tag{4.2}$$

Das Restglied $R_{n,q+1}$ ermittelt man durch Zusammenfassen der Restglieder bei Taylorentwicklung der einzelnen Summanden von $L y_n$ zu

$$R_{n,q+1} = \frac{h^{q+1}}{(q+1)!} \int_0^k K_q(\tau) \, y^{(q+1)}(x_n + h \cdot \tau) \, d\tau , \tag{4.3}$$

wobei der Kern für $j-1 \leq \tau \leq j$ $(j = 1,\ldots,k)$ durch

$$K_q(\tau) = \sum_{i \geq j} \left\{ a_i (i-\tau)^{q+1} - (q+1) \, b_i \, (i-\tau)^q \right\} \tag{4.4}$$

gegeben ist. Offenbar ist $K_q(\tau)$ in $0 \leq \tau \leq k$ erklärt, und es existieren dort stetige $(q-1)$-te Ableitungen $(q = 1,\ldots)$.

Der Operator L hat den genauen Grad $p \geq 1$, wenn $c_{p+1} \neq 0$,

$$\sum_{\varkappa=0}^{k} a_\varkappa = 0 \tag{4.5}$$

und

$$\sum_{\varkappa=0}^{k} (\varkappa^i \cdot a_\varkappa - i \cdot \varkappa^{i-1} \cdot b_\varkappa) = 0 \qquad \text{für } i = 1,\ldots,p \tag{4.6}$$

ausfallen. Man vergleiche hierzu (4.2).

---

1) Grundsätzlich werden alle einen Paragraphen betreffenden Voraussetzungen in den folgenden beibehalten, solange nichts anderes gesagt wird.

Es sei für $q = 1,2,\ldots$

$$\mathcal{L}_q = \left\{ L \mid p \geq q \right\} ;$$

offenbar gilt dann

$$\mathcal{L}_1 \supseteq \mathcal{L}_2 \supseteq \cdots$$

Wegen (1.6) wünscht man im allgemeinen, p möglichst groß zu machen. Wie G.Dahlquist 1956 gezeigt hat, ist es auch tatsächlich möglich, durch geeignete Wahl der Koeffizienten von L den Grad $p = 2k$ zu erreichen. Einen Operator dieses Grades erhält man, wenn man den höchsten Koeffizienten des Lagrange-Hermite-Interpolationspolynoms zu den Stützwerten

$$y_n, \ldots, y_{n+k} ; \; y'_n, \ldots, y'_{n+k}$$

gleich Null setzt. Wie Dahlquist aber weiter zeigt, ist die sich in einem großen p ausdrückende starke Angeglichenheit der Differenzengleichung an die Differentialgleichung mit einem Verlust an D-Stabilität gekoppelt. Für D-stabile Operatoren gilt nämlich

$$p \leq 2 \cdot \left[\frac{k+2}{2} \right] . \tag{4.7}$$

Mit Rücksicht hierauf und die von uns in Paragraph 3 erläuterte Zielsetzung soll im weiteren der Klasse $\mathcal{L}_{k+1}$ besondere Beachtung zuteil werden.

§ 5. Die Klasse $\mathcal{L}_{k+1}$.

Setzt man $\mathfrak{a} = (a_0, \ldots, a_k)$ und $\mathfrak{b} = (b_0, \ldots, b_k)$, so sind die Operatoren $L \in \mathcal{L}_{k+1}$ nach Paragraph 4 gekennzeichnet durch das Bestehen der Relationen (4.6), zusammenfaßbar zu

$$A\mathfrak{a} - B\mathfrak{b} = 0, \tag{5.1}$$

und die Nebenbedingung (4.5).

Die Matrizen A und B sind (k+1)-reihig-quadratisch. Die erste Spalte von A ist der Nullvektor, die erste von B der Einheitsvektor $\mathfrak{n}_1' = (1,0,\ldots,0)'$. Nimmt man jeweils die erste Spalte heraus, so sind die Restmatrizen Vandermondesche Matrizen. Man ersieht daraus, daß

$$\operatorname{Rang} A = k , \quad \operatorname{Rang} B = k+1 . \tag{5.2}$$

Somit ist durch (5.1) eine eindeutige Abbildung des vollen $\mathfrak{v}$- Raumes auf eine Hyperebene H des $\mathfrak{k}$ - Raumes gegeben, so daß es zu jedem Koeffizientensatz $\mathfrak{v}$, für den (4.5) erfüllt ist, genau einen Operator $L \in \mathfrak{L}_{k+1}$ mit diesen Koeffizienten gibt. Aus diesem Grunde soll $\varrho(s)$ das L erzeugende Polynom heißen. Die Hyperebene H wird unten genau bestimmt werden können.

Berücksichtigt man die Normierung $a_k = -1$, so erkennt man, daß die Operatoren $L \in \mathfrak{L}_{k+1}$ eine k-1 parametrige Schar bilden, die man durch geeignete Definitionen zu einem linearen Vektorraum machen kann. Darauf soll hier indes nicht näher eingegangen werden.

Die Beziehung (5.1) ist für numerische Zwecke ungeeignet. Die Auflösung des Systems linearer Gleichungen nach den b_ι führt zwangsläufig zu größeren Fehlern, die sich beim Gebrauch des Verfahrens L sehr unangenehm bemerkbar machen müssen.
Man kann die b_ι auch explizit angeben. Dazu bilde man eine Linearkomposition der Gleichungen (4.6) mit $p = k+1$, bei welcher alle $b_\varkappa$ mit $\varkappa \neq \iota$ herausfallen. Wegen der besonderen Gestalt der Koeffizienten in (4.6) läuft das auf die Lösung einer Interpolationsaufgabe hinaus. Man erhält für $0 \leq \iota \leq k$

(5.3) $$b_\iota = \sum_{\varkappa=0}^{k} \beta_{\iota\varkappa} \cdot a_\varkappa$$

mit

(5.4) $$\beta_{\iota\varkappa} = \int_0^\varkappa \ell_\iota(t)\,dt ,$$

$$\ell_\iota(t) = \prod_{\mu \neq \iota}^{0 \ldots k} \frac{t-\mu}{\iota-\mu} .$$

Das gleiche Verfahren, welches die b_ι lieferte, kann auch zu deren vollständigen Elimination aus (4.2) benutzt werden. Man findet so für $\iota = 1,2,\ldots$

(5.5) $$c_\iota = \iota \cdot \sum_{\varkappa=0}^{k} \gamma_{\iota\varkappa} \cdot a_\varkappa$$

mit

$$\gamma_{\iota\varkappa} = \int_0^\varkappa r_{\iota-1}(t)\,dt .$$

Dabei ist

$$r_i(t) = t^i - \varphi_i(t)$$

für $i = 0,1,\ldots$ das Restglied bei Lagrange-Interpolation der

Funktion t^i unter Verwendung der Stützstellen $t_\mu = \mu\,(\mu = 0,1,\ldots,k)$ und ist $\varphi_i = \varphi_i(t)$ das zugehörige Interpolationspolynom. Man hat also für $i = 1,2,\ldots$

$$r_{i-1}(t) = t^{i-1} - \sum_{\mu=0}^{k} \ell_\mu(t)\,\mu^{i-1},$$

woraus bei Integration sofort

$$\gamma_{i\varkappa} = \frac{1}{i}\varkappa^i - \sum_{\mu=0}^{k} \mu^{i-1}\cdot\beta_{\mu\varkappa} \tag{5.6}$$

folgt. Damit sind die ℓ_i und c_i mithilfe der $\beta_{\mu\varkappa}$ explizit durch die Koeffizienten des erzeugenden Polynoms ausgedrückt.

Von besonderem Interesse sind häufig die Taylorkoeffizienten c_{k+2} und c_{k+3}. Wie von der Lagrange-Interpolation her bekannt, ist

$$r_{k+1}(t) = \pi_k(t)\,,$$

wenn allgemein für $\varkappa = 0,1,\ldots$

$$\pi_\varkappa = \pi_\varkappa(t) = \prod_{\lambda=0}^{\varkappa}(t-\lambda)$$

gesetzt wird.

Versteht man nun, wenn $\varphi = \varphi(t)$ irgendein Polynom ist, unter $\bar{\varphi}$ dessen höchsten Koeffizienten, so gilt für die φ_i die mit $i = k+1$ beginnende Rekursion

$$\varphi_{i+1} = t\cdot\varphi_i - \bar{\varphi}_i\cdot\pi_k \qquad (i = k+1,\ldots)$$ [1)]

mit

$$\varphi_{k+1} = t^{k+1} - \pi_k(t)\,.$$

Wegen

$$\bar{\varphi}_{k+1} = \binom{k+1}{2}$$

erhält man somit nach einiger Umformung

$$r_{k+2} = \pi_{k+1} + \binom{k+2}{2}\pi_k\,,$$

und setzt man

$$\alpha_{\varkappa\nu} = \int_0^\nu \pi_\varkappa(t)\,dt \qquad \text{für } \varkappa,\nu = 0,1,\ldots,$$

so ergibt sich zuletzt aus (5.5)

$$c_{k+2} = (k+2)\sum_{\nu=0}^{k}\alpha_{k,\nu}\,a_\nu\,, \tag{5.7}$$

$$c_{k+3} = (k+3)\sum_{\nu=0}^{k}\alpha_{k+1,\nu}\,a_\nu + 3\cdot\binom{k+3}{3}\sum_{\nu=0}^{k}\alpha_{k,\nu}a_\nu\,. \tag{5.8}$$

---

1) $t\cdot\varphi_i$ interpoliert zwar schon t^{i+1} in der geforderten Weise, hat aber einen um 1 zu hohen Grad.

Die $\alpha_{\varkappa\nu}$ stehen in folgenden leicht herleitbaren Beziehungen:

Für $\varkappa = 2,4,\ldots$:

(5.9)
$$(\alpha)\quad \alpha_{\varkappa 0} = \alpha_{\varkappa\varkappa} = 0\,;$$
$$(\beta)\quad \alpha_{\varkappa\nu} = \alpha_{\varkappa,\varkappa-\nu}, \qquad \text{wenn } 0 \leqq \nu \leqq \varkappa;$$
$$(\gamma)\quad \alpha_{\varkappa,\varkappa+1} > \alpha_{\varkappa\nu} > 0, \quad \text{wenn } 1 \leqq \nu \leqq \varkappa-1.$$

Für $\varkappa = 1,3,\ldots$:

(5.10)
$$(\alpha)\quad \alpha_{\varkappa 0} = 0;$$
$$(\beta)\quad \alpha_{\varkappa\nu} + \alpha_{\varkappa,\varkappa-\nu} = \alpha_{\varkappa\varkappa} \text{ wenn } 0 \leqq \nu \leqq \varkappa\,;$$
$$(\gamma)\quad -\alpha_{\varkappa,\varkappa+1} < \alpha_{\varkappa\varkappa} < \alpha_{\varkappa 1} < 0\,;$$
$$(\delta)\quad \alpha_{\varkappa 1} < \alpha_{\varkappa\nu} < 0, \quad \text{wenn } 2 \leqq \nu \leqq \varkappa-1\,.$$

In den Relationen (β) drückt sich für $\varkappa \equiv 0\,(2)$ die Ungeradheit, für $\varkappa \equiv 1\,(2)$ die Geradheit der Funktion $\pi_\varkappa\left(\frac{\varkappa}{2}+u\right)$ aus. Auf eine wichtige Folge dieser Eigenschaften kommen wir in § 10 zurück.

Wegen (5.3) ist es nunmehr möglich, die Hyperebene H, in welche der volle m-Raum vermöge (5.1) abgebildet wird, genau zu bestimmen. Es folgt nämlich aus (5.3)

$$\sum_{\iota=0}^{k} (-1)^{k-\iota}\cdot \iota!\cdot(k-\iota)!\cdot b_\iota = \sum_{\varkappa=0}^{k} a_\varkappa \int_0^\varkappa \sum_{\iota=0}^{k} \prod_{\lambda\neq\iota} (t-\lambda)\,dt = \sum_{\varkappa=0}^{k} a_\varkappa \int_0^\varkappa \pi_k'(t)\,dt\,.$$

Wegen $\pi_k(\varkappa) = 0$ für $\varkappa = 0,1,\ldots,k$ erhält man hieraus für H die Darstellung

$$H: \qquad \sum_{\iota=0}^{k} (-1)^{k-\iota}\,\iota!\,(k-\iota)!\,b_\iota = 0$$

§6. Die Klasse $\mathcal{L}_k$ (offene Operatoren).

Die Formel (5.3) gestattet uns, von irgendeinem Koeffizientensatz m ausgehend, einen Operator $L \in \mathcal{L}_{k+1}$ auf numerisch einfache Art zu berechnen. Im allgemeinen wird der so erzeugte Operator geschlossen sein, d.h. es ist $b_k \neq 0$. Geschlossene Operatoren kann man nur zur Verbesserung von Schätzungen für y_{n+k} benutzen, während man für die Schätzung selbst einen offenen Operator L^* ($b_k = 0$) zusätzlich heranziehen muß. Solche offenen L^* sind in $\mathcal{L}_k$ enthalten.

Die Klasse $\mathfrak{L}_k$ ist gekennzeichnet durch die für $p = k$ genommenen Gleichungen (4.5) und (4.6). Auf demselben Wege, auf dem in § 5 die Formel (5.3) gewonnen wurde, können wir jetzt - wenn etwa b_k nicht eliminiert, sondern als Parameter beibehalten wird - für $\iota = 0,1,\ldots,k-1$ zu folgenden Darstellungen gelangen:

$$(6.1) \qquad b_\iota = \sum_{\varkappa=0}^{k} \beta^*_{\iota\varkappa} \cdot a_\varkappa + (-1)^{k-\iota} \cdot \left\{ \begin{matrix} k \\ \iota \end{matrix} \right\} \cdot b_k$$

mit

$$\beta^*_{\iota\varkappa} = \int_0^\varkappa \ell^*_\iota(t)\,dt\,,$$

$$\ell^*_\iota(t) = \prod_{\mu \neq \iota}^{0,\ldots,k-1} \frac{t-\mu}{\iota-\mu}\,.$$

Die $\beta^*_{\iota\varkappa}$ stimmen also mit den für k-1 statt k genommenen $\beta_{\iota\varkappa}$ überein.

Die offenen L^* von einem Grade $p \geq k$ sind nach (6.1) eindeutig gekennzeichnet durch das Bestehen der Beziehung $b_\iota = b^*_\iota$ mit

$$(6.1)a \qquad b^*_\iota = \sum_{\varkappa=\iota}^{k} \beta^*_{\iota\varkappa} \cdot a_\varkappa \quad \text{für } \iota = 0,\ldots,k-1.$$

Aus (4.2) folgt für diese Operatoren ferner mit $c_{k+1} = c^*_{k+1}$

$$c^*_{k+1} = \sum_{\iota=0}^{k} \left(\iota^{k+1} \cdot a_\iota - (k+1)\,\iota^k \cdot b^*_\iota \right)$$

oder wegen $\iota^{k+1} = (k+1) \cdot \int_0^\iota t^k\,dt$ und (6.1)a

$$c^*_{k+1} = (k+1) \left\{ \sum_{\iota=0}^{k} a_\iota \int_0^\iota t^k\,dt - \sum_{\iota=0}^{k} \iota^k \sum_{\varkappa=0}^{k} \beta^*_{\iota\varkappa} a_\varkappa \right\}.$$

Hieraus folgt bei Berücksichtigung der Definition für $\beta^*_{\iota\varkappa}$

$$c^*_{k+1} = (k+1) \sum_{\varkappa=0}^{k} a_\varkappa \int_0^\varkappa \left\{ t^k - \sum_{\iota=0}^{k} \ell^*_\iota(t) \cdot \iota^k \right\} dt.$$

Der Integrand ist bestimmt als das Restglied bei Lagrange-Interpolation der Funktion t^k mittels der Stützstellen $t_\mu = \mu$, $\mu = 0,..,k-1$. Es ist also, wie bekannt,

$$t^k - \sum_{\iota=0}^{k} \ell^*_\iota(t) \cdot \iota^k = \pi_{k-1}\,,$$

und man erhält

$$(6.2) \qquad c^*_{k+1} = (k+1) \cdot \sum_{\varkappa=0}^{k} \alpha_{k-1,\varkappa} \cdot a_\varkappa\,.$$

Wird somit $\varrho(s)$ irgendwie so vorgegeben, daß $\varrho(1) = 0$ ausfällt, so sind dadurch zwei Operatoren eindeutig bestimmt:

$L \in \mathcal{L}_{k+1}$ vermöge (5.3),

$L^* \in \mathcal{L}_k$ vermöge (6.1),

wobei der letztere offen ist.

Nach (5.3) ist

$$k! \cdot b_k = k! \sum_{\varkappa=0}^{k} a_\varkappa \int_0^\infty \ell_k(t)\,dt = \sum_{\varkappa=0}^{k} a_\varkappa \int_0^\infty \pi_{k-1}(t)\,dt,$$

und aus (6.2) folgt hiermit

$$c^*_{k+1} = (k+1)! \cdot b_k \quad . \tag{6.3}$$

Jedem geschlossenen $L \in \mathcal{L}_{k+1}$ entspricht somit vermöge der Abbildung $L \leftrightarrow L^*$ genau ein offener Operator $L^* \in \mathcal{L}_k$, der den genauen Grad $p = k$ besitzt.

Ferner erkennt man an (6.3): Ist L "stark geschlossen", d.h. ist $|b_k|$ groß, so hat L^* im allgemeinen einen betragsmäßig großen Formelfehler t_n, so daß L^* schlechte Schätzungen für y_{n+k} liefern wird. Zugleich hat das Verfahren (1.3) eine große Lipschitz-Konstante (vgl. § 9, S. 20), was ein langsames Laufen der Verbesserungsroutine zufolge hat. Mithin muß man, wenn man nur die Rechengeschwindigkeit und nicht zugleich die erzielte Genauigkeit im Auge hat, bestrebt sein, $|b_k|$ möglichst klein zu halten.

Aufgrund der Darstellungen der letzten beiden Paragraphen sind wir u.a. in der Lage, geschlossene und offene Operatoren auf numerisch einfache Art zu erzeugen, vorausgesetzt, daß die $\beta_{i\varkappa}$ ein für allemal berechnet sind. Es darf hier angemerkt werden, daß diese Matrizen für $k \leq 12$ mithilfe der elektronischen Rechenanlage Siemens 2002 am Rechenzentrum in Tübingen auf zehn Dezimalstellen genau ermittelt und in ALGOL-Norm in Lochstreifen gestanzt wurden.

K A P I T E L III

Absolute Stabilität.

§7. Vorbemerkungen.

Nach H.Rutishauser 1952 sind für die Instabilität eines Verfahrens L die eingeschleppten Nebenlösungen der Differenzengleichung (1.4) verantwortlich. Man könnte nun meinen, daß man das Verfahren mit dem günstigsten Fehlerverlauf dann erhält, wenn man $s_2=\ldots=s_k=0$ macht (Adams'sches Verfahren), um so die Nebenlösungen gewaltsam zu unterdrücken bzw. um dafür zu sorgen, daß diese zumindest rasch abklingen. Wie man jedoch schon an einem Beispiel im Falle $k=2$ erkennt, ist diese Vermutung nicht richtig.

Im Falle $k=2$ haben die Verfahren $L \in \mathfrak{L}_{k+1}$ die Darstellung

$$(7.1) \qquad u_{n+2} = (1-a_1)\, u_n + a_1 u_{n+1} + \frac{h}{12}\Big\{(4-5a_1)\, U_n + 8(2-a_1)\, U_{n+1} + (4+a_1)\, U_{n+2}\Big\}.$$

Es ist $s_2 = 1-a_1$; das Verfahren (7.1) ist also D-stabil für

$$0 \leq a_1 < 2 .$$

Für $a_1 = 0$ ist in (7.1) die Simpson-Regel, für $a_1 = 1$ das Adams'sche Verfahren 2.Ordnung enthalten.

Es sei $\eta_n = u_n - y_n$. Betrachtet werde das Integrationsbeispiel

$$y' = -y\,, \quad y(0) = 1\,;$$
$$\eta_0 = K\,, \qquad \eta_1 = -K \quad \text{mit } K = 10^{-4}\,; \quad h = 0.2\,.$$

Runde- und Abbrechfehler sollen dabei dem Betrage nach kleiner als 10^{-6} gehalten werden, so daß sie keinen wesentlichen Einfluß auf den Verlauf von η_n für $n=2,\ldots$ haben. In Tabelle I ist das Ergebnis der Rechnung angeführt, und zwar ist die Funktion

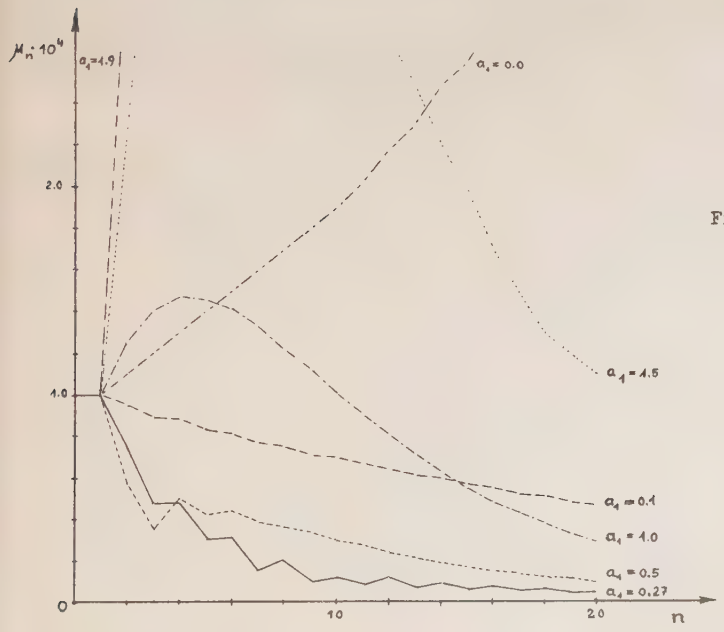
$$\mu_n = \max_{i=-1,0,1} |\eta_{n+i}| \text{ für } n = 2,3,\ldots\,, \quad \mu_0 = \eta_0\,, \quad \mu_1 = \eta_1$$

für verschiedene Werte von a_1 aus dem Bereich der D-Stabilität tabelliert, welche gegenüber η_n einen glatteren Verlauf hat. Eine anschauliche Vorstellung von μ_n vermittelt Fig.1 .

n \ a_1	0.00	0.10	0.27	0.50	1.00	1.50	1.90
0	1.0	1.00	1.00	1.00	1.00	1.0	1.0
1	1.0	1.00	1.00	1.00	1.00	1.0	1.0
2	1.1	.95	.74	.57	1.24	2.3	3.4
3	1.2	.89	.48	.35	1.4o	3.3	5.8
4	1.3	.88	.48	.50	1.47	3.9	8.0
5	1.4	.83	.30	.42	1.46	4.2	10.2
6	1.5	.81	.31	.44	1.41	4.3	12.1
7	1.6	.77	.15	.39	1.33	4.2	13.4
8	1.7	.75	.20	.36	1.23	4.0	14.5
9	1.8	.71	.10	.34	1.13	3.7	15.4
10	1.9	.70	.07	.30	1.02	3.4	15.8
11	2.0	.66	.08	.27	.91	3.1	16.0
12	2.2	.64	.12	.24	.81	2.8	16.1
13	2.3	.61	.07	.21	.72	2.5	15.9
14	2.5	.60	.09	.19	.63	2.2	15.6
15	2.6	.56	.06	.17	.55	2.0	15.1
16	2.8	.55	.07	.15	.49	1.7	14.5
17	3.0	.52	.05	.14	.43	1.5	13.7
18	3.2	.51	.06	.12	.38	1.3	13.3
19	3.5	.48	.04	.11	.33	1.2	12.6
20	3.7	.47	.05	.09	.29	1.1	11.9

Tafel I

($\mu_n \cdot 10^4$)



Figur 1

Man sieht ohne weiteres, daß μ_n bei den Verfahren mit $1 \leq a_1 < 2$ anfänglich einen sehr viel ungünstigeren Verlauf aufweist, als bei den übrigen.
Zwar ist μ_n für große n im Falle $a_1 = 1$ offensichtlich kleiner als als z.B. für $a_1 = 0.1$, jedoch muß damit gerechnet werden, daß im realen Fall, in dem Runde- und Abbrechfehler nicht vernachlässigt werden können, lokale Störungen etwa bei $n = 2,3$ oder 4 den von uns fixierten Fehlerverlauf längst zunichte gemacht und für ein weiteres Anschwellen von μ_n gesorgt haben. Anders ausgedrückt: lokale Störungen bei $n = 2,3,\ldots$ können dafür sorgen, daß die Koeffizienten A_i in (2.2) sich von Schritt zu Schritt ändern, und zwar so, daß μ_n sich trotz raschen Abklingens der s_i^n für $n \to \infty$ $(i=2,\ldots,k)$ immer weiter hochschaukelt.

Aus dem oben Gesagten geht hervor, daß es sinnvoller ist, das lokale Verhalten von μ_n zur Beurteilung eines Verfahrens heranzuziehen als asymptotische für $n \to \infty$. Eine Klassifikation der D-stabilen Operatoren L nach ihrem lokalen Verhalten bewirkt der Begriff der absoluten Stabilität, der von uns - nach Klärung einiger Grundbegriffe in Paragraph 8 - nunmehr eingeführt wird.

§8. Die Variationsgleichung.

Das Integral y(x) von (1.1), genauer: $y_n = y(x_n)$ genügt der Differenzengleichung

(8.1) $$L y_n = t_n \,.$$

Das Differenzenverfahren L besteht darin, daß statt (8.1)

(8.2) $$L u_n = 0 \qquad (n = 0,1,\ldots)$$

gelöst wird. Auch dies geschieht mit einem Fehler; statt u_n erhält man eine Zahlenfolge $\tilde{y}_n$, für welche

(8.3) $$L \tilde{y}_n = r_n$$

ausfällt. Wir nennen r_n den **Rundefehler** beim Schritt n. Er setzt sich zusammen aus einer statistischen Komponente, dem eigentlichen Rundungsfehler, und einer systematischen (nur bei Verwendung geschlossener Operatoren), dem Abbrechfehler bei iterativer

Auflösung der Differenzengleichung (8.2) nach u_{n+k}. Außer dem Formelfehler t_n und dem Rundefehler r_n bei jedem vorangegangenen Schritt haben noch die Anfangswertfehler $\eta_0,\dots,\eta_{k-1}$ Einfluß auf den Gesamtfehler

(8.4) $$\eta_n = \tilde{y}_n - y_n \qquad (n = 0,\dots,N = \left[\frac{b-a}{h}\right]).$$

Die Bezeichnung ist so gewählt, daß der Gesamtfehler für n=0,...,k-1 mit dem Anfangswertfehler übereinstimmt.

Subtrahiert man (8.1) von (8.3), so erhält man bei Berücksichtigung von (8.4) die Differenzen-Variationsgleichung

(8.5) $$\sum_{\varkappa=0}^{k}\left\{a_\varkappa - h\,b_\varkappa\, g^*_{n+\varkappa}\right\}\eta_{n+\varkappa} = r_n - t_n$$

(vgl. hierzu (1.2)). Es ist dabei

$$g^*_{n+\varkappa} = f_y(x_{n+\varkappa}, y^*_{n+\varkappa})$$

zu setzen mit zwischen $y_{n+\varkappa}$ und $y_{n+\varkappa} + \eta_{n+\varkappa}$ gelegenen Werten $y^*_{n+\varkappa}$ und von vornherein vorauszusetzen, daß die Verbindungsstrecken der Punkte (x_m, y_m) und $(x_m, \tilde{y}_m)$ für $m = 0,\dots,N$ ganz in dem Gebiete G enthalten sind.

§9. Definitionen. Kriterien für absolute Stabilität.

Definition 1: Eine Differentialgleichung (1.1) heißt stabil oder zur Klasse $\mathfrak{S}_r$ gehörig, wenn $f \in \mathfrak{F}_r$, $r \geq 1$, und wenn es ein Teilgebiet $G^*(x,y) = G^* \subseteq G$ gibt, so daß die Integralkurve $(x, y(x))$ für $a \leq x \leq b$ ganz in G^* enthalten ist und so daß

$$f_y(x,y) < 0 \text{ in } G^*(x,y).$$

Wir setzen für $p \geq 1$

$$\mathfrak{K}_p = \left\{ L \mid L \in \mathfrak{L}_p,\ a_\varkappa > 0 \text{ für } 0 \leq \varkappa \leq k-1 \right\}^{1)},$$

$$\overline{\mathfrak{K}}_p = \left\{ L \mid L \in \mathfrak{L}_p,\ a_\varkappa \geq 0 \text{ für } 0 \leq \varkappa \leq k-1 \right\}.$$

Offenbar gilt dann

$$\mathfrak{K}_1 \supseteq \mathfrak{K}_2 \supseteq \dots\ ,\quad \overline{\mathfrak{K}}_1 \supseteq \overline{\mathfrak{K}}_2 \supseteq \dots\ .$$

---

1) Man beachte (1.5).

Definition 2 : Ein Operator L heißt positiv (nichtnegativ), wenn er zur Klasse $\mathfrak{P}_1$ ($\overline{\mathfrak{P}}_1$) gehört.

Definition 3 : Ein Operator L heißt in Bezug auf eine Klasse von Differentialgleichungen (1.1) absolut stabil, wenn es zu jeder festen Differentialgleichung der Klasse ein $K_0 > 0$ gibt, so daß zu jedem K: $0 < K \leq K_0$ ein $h(K) > 0$ existiert, derart daß für alle h mit

$$0 < h \leq h(K)$$

aus dem Bestehen der Ungleichungen

(9.1) $|\eta_x| \leq K \quad (0 \leq x \leq k-1)$

die Abschätzung

(9.2) $|\eta_n| \leq K \quad (0 \leq n \leq N)$

folgt, wenn nur die Rundefehler einer Bedingung

(9.3) $|r_n| \leq r(h)$

genügen, bei der r(h) eine in $0 < h \leq h(K)$ erklärte und positive Funktion bedeutet, für die der folgende Grenzwert existiert und den angegebenen Wert hat:

(9.4) $\lim\limits_{h \to 0} r(h)/h = 0.$

Definition 4 : Ein Operator L heißt absolut stabil, wenn $p \geq 1$ ist und L in Bezug auf $\mathfrak{P}_2$ absolut stabil ist.

Wir wenden uns nun der Frage zu, welche Operatoren L absolut stabil sind. Hierüber gilt zunächst der

Satz 1 (notwendige Stabilitätsbedingung)
Ein Operator L ist höchstens dann absolut stabil, wenn er der Klasse $\overline{\mathfrak{P}}_1$ angehört.

Beweis: Wir nehmen an, L sei absolut stabil und es sei

$$a_\iota < 0 \quad \text{für ein } \iota : 0 \leq \iota \leq k-1.$$

Wegen $p \geq 1$ und (4.5) ist dann

(9.5) $$\sum_{\varkappa=0}^{k-1} |a_\varkappa| > 1.$$

Wir betrachten die Differentialgleichung

$$y' = g\,y\,, \qquad y(0) = 0\,, \qquad 0 \leq x \leq 1\,.$$

Hier sei $g < 0$ eine Konstante. Offenbar ist diese Differentialgleichung in $\mathfrak{F}_2$ enthalten.

Es sei $K > 0$ irgendwie vorgegeben. Die Anfangswerte seien

$$\tilde{y}_\varkappa = K \cdot \operatorname{sgn} a_\varkappa \qquad (0 \leq \varkappa \leq k-1)\ ^{1)}.$$

Außerdem werde $r_n \equiv 0$ vorausgesetzt. Dann sind (9.1) - wegen $y_0 = \ldots = y_{k-1} = 0$ - und eine Bedingung (9.3) sicher erfüllt.

Die Variationsgleichung (8.5) kann gebildet werden, und man erhält daraus für $n = 0$ nach einfacher Umformung

$$(9.6) \qquad (1 + H b_k)\eta_k = \sum_{\varkappa=0}^{k-1}{}' |a_\varkappa - H b_\varkappa| \cdot K + t_0$$

mit $H = g\,h$. Bei der Summenbildung sind dabei die $\varkappa$ auszulassen, für welche $a_\varkappa = 0$ ist. Wegen $H \to 0$ und $t_0 \to 0$ - das letztere folgt aus p 1 - erhält man aus (9.6) und (9.5) für $h \to 0$

$$\eta_k \longrightarrow \sum_{\varkappa=0}^{k-1}{}' |a_\varkappa| \cdot K = K \cdot \sum_{\varkappa=0}^{k-1}{}' |a_\varkappa| > K\,;$$

das heißt aber, daß (9.2) für alle hinreichend kleinen $h > 0$ bereits bei $n = k$ verletzt ist. Daher ist L nicht absolut stabil. Die Annahme $a_k < 0$ ist also falsch und der Satz bewiesen.

Satz 2 (hinreichende Stabilitätsbedingung)

Alle Operatoren L der Klasse $\mathfrak{P}_1$ sind absolut stabil.

Beweis: Es sei $L \in \mathfrak{P}_1$ und die Differentialgleichung (1.1) zu $\mathfrak{F}_2$ gehörig.

Es gibt dann eine Konstante $K_1 > 0$, so daß die kompakte Punktmenge

$$\mathfrak{F} = \left\{ (x,y) \;\middle|\; a \leq x \leq b,\ |y - y(x)| \leq 5K_1 \right\}$$

ganz in dem Gebiet G^* (vgl. Definition 1) enthalten ist. Es gilt dann mit einer geeigneten Konstanten g

$$(9.7) \qquad f_y(x,y) \leq g < 0 \quad \text{in } \mathfrak{F}\,.$$

Möglicherweise ist G(x,y) die ganze (x,y)-Ebene R_2 und gilt

$$(9.8) \qquad |f|\,,\ |f_y|\,,\ |f_{yx}|\,,\ |f_{yy}| \leq A \quad \text{in } R_2\,.$$

Ist das nicht richtig, so können wir f nach einem von E.Kamke 1961 [2)]

---

1) sgn x = +1,0 oder -1, je nachdem ob x >, = oder < 0 ist.
2) E.Kamke 1961 Bd.II S.24.

benutzten Verfahren so zu einer Funktion f^* abändern, daß diese in der ganzen Ebene R_2 erklärt ist und stetige und beschränkte partielle Ableitungen bis zur zweiten Ordnung besitzt und so daß f^* in $\mathfrak{F}$ mit f übereinstimmt. Die weiteren Erörterungen können dann mit f^* statt f angestellt werden.

Ist L offen, so ist durch die Differenzengleichung (1.4) stets ein lineares Differenzenverfahren (1.3) bestimmt. Wenn L aber geschlossen ist, so muß dieses erst nachgewiesen werden. Offenbar ist es möglich, die Differenzengleichung (1.4) auf die Gestalt

$$u_{n+k} = F_n(u_{n+k})$$

zu bringen (vgl.(1.2)), wobei $F_n(u)$ einer Lipschitzbedingung mit der Lipschitzkonstanten $\lambda = h \cdot A \cdot |b_k|$ genügt. Ist $h > 0$ nur hinreichend klein, etwa $0 < h < h_3$, so gilt $\lambda < 1$, und setzt man, von irgendeinem Wert $\overset{(0)}{u}_{n+k}$ ausgehend, die Iteration

$$\overset{(i+1)}{u}_{n+k} = F_n(\overset{(i)}{u}_{n+k}) \qquad (i = 0,1,\ldots)$$

an, so weiß man, daß die Folge der $\overset{(i)}{u}_{n+k}$ gegen eine Lösung u_{n+k} der Differenzengleichung (1.4) konvergiert, das heißt, es gibt eine Lösung $u_{n+k} = \Lambda(u_n,\ldots,u_{n+k-1})$ - vgl.(1.3). Diese Lösung kann wohl vom Ausgangswert $\overset{(0)}{u}_{n+k}$ und damit vom verwendeten offenen Operator abhängen, darauf kommt es hier jedoch nicht an. Ausgehend von irgendwelchen Anfangswerten $\tilde{y}_\varkappa$ $(0 \leq \varkappa \leq k-1)$ erhält man jedenfalls eine wohldefinierte Lösung $\tilde{y}_n$ $(n = 0,\ldots,N)$, und es kann die Variationsgleichung (8.5) gebildet werden.

Mit $q = 1$ erhält man aus (4.1) und (4.3) unter Berücksichtigung der Beschränktheit von $y''(x)$ in $a \leq x \leq b$ und der Beschränktheit von $K_1(\tau)$ in $0 \leq \tau \leq k$ für $|t_n|$ eine Schranke:

$$|t_n| \leq t(h) = Mh^2 \tag{9.9}$$

mit einer geeigneten Konstanten $M > 0$. Da L positiv ist, kann man ferner ein h_2 mit $0 < h_2 < h_3$ bestimmen, so daß

$$\begin{aligned} A\,h\,|b_\varkappa| &\leq a_\varkappa \qquad (0 \leq \varkappa \leq k-1) \\ A\,h\,|b_k| &\leq \frac{1}{2} \end{aligned} \tag{9.10}$$

für $0 < h \leq h_2$ gilt.

Im weiteren sei stets $0 < h \leq h_1$, $0 < K \leq K_1$, und für die Rundefehler gelte eine Beziehung (9.3).
Wir setzen voraus, daß die Bedingung

(9.11) $\qquad |\eta_\nu| \leq K \quad \text{für } 0 \leq \nu \leq n+k-1$

für ein $n \geq 0$ bereits erfüllt sei. Für $n = 0$ ist das der Fall, wenn (9.1) vorausgesetzt wird, was nunmehr geschehe. Für $n = N-k+1$ stimmt (9.11) mit (9.2), also der Behauptung, überein. Es ist zu zeigen, daß es ein $K_0 > 0$ und eine Funktion $h(K) > 0$ gibt, so daß für $K \leq K_0$ und $h \leq h(K)$ aus (9.11) die Abschätzung $|\eta_{n+k}| \leq K$ folgt, und zwar unabhängig von n. Ein Induktionsschluß vollendet dann den Beweis.

Für $0 < h \leq h_2$, $0 < K < K_1$ folgt aus (8.5)

(9.12) $\qquad (1 + h\, b_k\, g^*_{n+k})\, |\eta_{n+k}| \leq \sum_{\varkappa=0}^{k-1} (a_\varkappa - h\, b_\varkappa\, g^*_{n+\varkappa})\, K + t(h) + r(h)\,,$

und hieraus erhält man die (sehr grobe) Abschätzung

$$|\eta_{n+k}| \leq Y(K,h)$$

mit

$$Y(K,h) = 2\Big\{\sum_{\varkappa=0}^{k-1} 2\,a_\varkappa K + t(h) + r(h)\Big\}$$
$$= 4K + 2\,t(h) + 2\,r(h)$$

nach (9.10) und (4.5). Offenbar gibt es also ein h_1 mit $0 < h_1 \leq h_2$, so daß

(9.13) $\qquad Y(K,h) \leq 5K_1 \qquad \text{für } 0 < h \leq h_1;$

und es gilt

(9.14) $\qquad \lim_{K,h \to 0} Y(K,h) = 0\,.$

Aus (9.13) folgt, daß die wegen (9.11) ohnehin für $\nu = 0,\ldots,n+k-1$ schon geltende Relation

$$(x_\nu, y^*_\nu) \in \mathfrak{F}$$

auch für $\nu = n+k$ erfüllt ist. Setzt man also allgemein $g_\nu = f_y(x_\nu, y_\nu)$ für $\nu = 0,\ldots,N$, so findet man wegen (9.8)

(9.15) $\qquad |g_{n+\varkappa} - g_n| \leq A\,(\varkappa h + |y_{n+\varkappa} - y_n|) \leq A\,(1+A)\,kh,$

und außerdem gilt dann

(9.16) $\qquad |g^*_{n+\varkappa} - g_{n+\varkappa}| \leq A\,|\eta_{n+\varkappa}| \leq \begin{cases} A \cdot K & \text{für } 0 \leq \varkappa \leq k-1 \\ A \cdot Y(K,h) & \text{für } \varkappa = k \end{cases}.$

Nunmehr können wir wegen (9.15), (9.16) und (7), (9.14) zwei Zahlen h_0 und K_0 mit $0 < h_0 \leq h_1$, $0 < K_0 \leq K_1$ bestimmen, so daß für alle h, K mit $0 < h \leq h_0$, $0 < K \leq K_0$

$$\sum_{x=0}^{k} |b_x| \cdot |g^*_{n+x} - g_n| \leq \frac{1}{2}|g| \tag{9.17}$$

ausfällt.

Damit nähern wir uns dem Ziel. Es sei weiterhin

$$0 < h \leq h_0 , \quad 0 < K \leq K_0 .$$

Wegen (9.12) und (9.10) ist $|\eta_{n+k}| \leq K$ sicher immer dann, wenn

$$\sum_{x=0}^{k-1} (a_x - h\, b_x\, g^*_{n+x}) \cdot K + t(h) + r(h) \leq (1 + h\, b_k\, g^*_{n+k}) \cdot K$$

oder - wegen (4.5) - wenn

$$\left\{ t(h) + r(h) \right\} \cdot h^{-1} \leq K \sum_{x=0}^{k} b_x\, g^*_{n+x} . \tag{9.18}$$

Für nichtnegative L ist nach (4.5) und (4.6)

$$\sum_{x=0}^{k} b_x = \sum_{x=0}^{k} x\, a \leq -k + (k-1) \sum_{x=0}^{k-1} a_x = -1 , \tag{9.19}$$

und man erhält aus der Identität

$$\sum_{x=0}^{k} b_x\, g^*_{n+x} = g_n \sum_{x=0}^{k} b_x + \sum_{x=0}^{k} b_x (g^*_{n+x} - g_n)$$

mihilfe von (9.17) und (9.19) wegen $g_n \leq g < 0$ (vgl. (9.7))

$$K \sum_{x=0}^{k} b_x\, g^*_{n+x} \geq K \left(|g| - \frac{1}{2}|g| \right) = \frac{1}{2} K |g| > 0 . \tag{9.20}$$

Aus (9.20) und (9.4), (9.9) folgt unmittelbar, daß es zu jedem K mit $0 < K \leq K_0$ ein $h(K) > 0$ gibt, so daß (9.18) für alle festen h mit $0 < h \leq h(K)$ erfüllt ist, mithin, so daß aus (9.11) $|\eta_{n+k}| \leq K$ folgt. Die Bestimmung von h_0 und K_0 geschah unabhängig von n.

Damit ist der Beweis beendet für den Fall, daß $G = R_2$ war und f nicht zu einer Funktion f^* abgeändert werden mußte. In diesem letzteren Falle stimmen nicht nur die Integrale der mit f und mit f^* gebildeten Differentialgleichungen überein, sondern auch die numerischen Lösungen der Differenzengleichungen, vorausgesetzt, daß $K \leq K_0$ und $h \leq h(K)$ genommen werden. Die absolute Stabilität von L in Bezug auf die mit f gebildete Differentialgleichung folgt somit unmittelbar aus der bezüglich f^*, so daß alles bewiesen ist.

Anmerkungen:

1) Die Forderung (9.4) bezüglich der Rundefehler ist offenbar eine Minimalforderung. In der Praxis wird man eher $r(h) \approx t(h)$ haben wollen, wenn man bei Verkleinerung der Schrittweite einen bedeutenden Genauigkeitsgewinn bzw. bei vorgegebener Fehlerschranke eine große zulässige Schrittweite erzielen will. Für den Begriff der absoluten Stabilität jedoch ist es unwesentlich, ob man (9.4) durch eine schärfere Bedingung ersetzt oder nicht.

2) Es ist prinzipiell unmöglich, in Definition 3 statt der Schranke K für den absoluten Fehler mit einer solchen für den relativen Fehler auszukommen, wenn man erreichen will, daß es einen in Bezug auf eine stabile lineare Differentialgleichung absolut stabilen Operator L gibt. Ganz abgesehen davon würden Schwierigkeiten bei identisch verschwindenden Integralen entstehen.

3) Die Auszeichnung der stabilen Differentialgleichungen in Definition 4 kann nicht vermieden werden. Man kann nämlich zeigen, daß kein Operator L in Bezug auf eine instabile Differentialgleichung absolut stabil ist.

4) Die Forderung in Definition 3, daß (9.2) aus (9.1) für a l l e $h \leq h(K)$ folgen solle, scheint zunächst reichlich stark zu sein. Man könnte auch zufrieden sein, wenn dies für e i n $h > 0$ gilt (so W.Liniger 1957). Diese Schrittweite sollte man dann aber auch genau kennen. - Der Begriff der absoluten Stabilität ist so beschaffen, daß man bei Verkleinerung der Schrittweite zwar möglicherweise unnötige Rechenarbeit, niemals aber Verlust an Stabilität in Kauf nehmen muß.

Die Frage, welche Operatoren $L \in \overline{\mathfrak{R}}_1$, für die $a_\iota = 0$ für ein ι mit $0 \leq \iota \leq k-1$ ausfällt, absolut stabil sind, ist noch nicht allgemein beantwortet. Sicher ist $L \in \overline{\mathfrak{R}}_1$ absolut stabil, wenn mit a_ι auch b_ι verschwindet. Das geht bei geringfügiger Modifikation des Beweises zu Satz 2 hervor. Dagegen ist die Simpson-Regel ($\varrho(s) = -s^2 + 1$, $p = 4$)

sicher nicht absolut stabil. Das geht hervor aus

Satz 3: Ist k gerade, so ist der durch $\varrho(s) = -s^k + 1$ erzeugte Operator $L \in \bar{\mathcal{R}}_{k+1}$ nicht absolut stabil.

Beweis: Es sei k gerade, $\varrho(s) = -s^k + 1$ und $p \geq k+1$. Nach (5.3) ist dann

$$b_0 = -\frac{1}{k!}\int_0^k \pi_{k-1}(t-1)\,dt = -\frac{1}{k!}\int_0^k \pi_{k-1}(k-1-\tau)\,d\tau = -\frac{1}{k!}\int_0^k \pi_{k-1}(\tau)\,d\tau = b_k .$$

Außerdem ist nach (5.10), (*)

$$b_k = -\frac{1}{k!}\,\alpha_{k-1,k} > 0 .$$

Für die stabile Differentialgleichung $y' = -y$ und $r_n \equiv 0$, $\eta_1 = \ldots = \eta_{k-1} = 0$ nimmt somit die Variationsgleichung (8.5) für $n = 0$ nach einfacher Umformung die Gestalt

$$\eta_k = \frac{1 + h\,b_k}{1 - h\,b_k}\,\eta_0 + O(h^3)$$

an. Ist $\eta_0 = K > 0$, so hat man offenbar $\eta_k > K$, wenn h nur hinreichend klein genommen wird. Das heißt aber, daß L nicht absolut stabil ist.

Wir wollen jetzt zeigen, daß der Begriff der absoluten Stabilität stärker ist als der der D-Stabilität.

Satz 4: Jeder absolut stabile Operator L ist D-stabil. Die Umkehrung gilt nicht.

Beweis: Daß die Umkehrung nicht gilt, folgt aus Satz 3. Ferner genügt es wegen Satz 1, die D-Stabilität für alle Operatoren $L \in \bar{\mathcal{R}}_1$ nachzuweisen.

Es sei $L \in \bar{\mathcal{R}}_1$ und $\varrho(s)$ das L erzeugende Polynom. Spaltet man von $\varrho(s)$ den wegen (4.5) vorhandenen Linearfaktor $(1 - s)$ und die größt-mögliche Potenz s^r ab, so erhält man

$$\varrho(s) = (1 - s)\cdot s^r \cdot \varrho^*(s)$$

mit $0 \leq r \leq k-1$ und

$$\varrho^*(s) = \sum_{\varkappa=r}^{k-1} A^*_\varkappa s^{\varkappa - r} ,$$

$$A^*_\varkappa = a_0 + \ldots + a_\varkappa , \quad A^*_r > 0 .$$

Ist $r = k-1$, so ist $\varrho^*(s) \equiv 1$ und L offenbar D-stabil (vgl. die Bedingung (D) von S.4). Ist aber $k \geq 2$ und $r \leq k-2$, so gilt

$$0 < A^*_r \leq A^*_{r+1} \leq \ldots \leq A^*_{k-1} = 1 ,$$

und es gilt für jede Wurzel s^* von $\varphi^*(s)$ die bekannte Einschließung

$$0 < \min_{r+1 \leq \varkappa \leq k-1} \frac{A_{\varkappa-1}}{A_\varkappa} \leq |s^*| \leq \max_{r+1 \leq \varkappa \leq k-1} \frac{A_{\varkappa-1}}{A_\varkappa} \leq 1 .$$

Mithin ist nur noch zu prüfen, ob $\varphi^*(1) = 0$ ist oder $\varphi^*(s)$ eine mehrfache Wurzel s^* mit $|s^*| = 1$ hat.

Der erste Fall kann gleich ausgeschlossen werden, denn es ist

$$\varphi^*(1) = \sum_{\varkappa=r}^{k-1} A^*_\varkappa \geq A_r > 0 .$$

Der zweite Fall kann nur unter der Bedingung $k \geq 3$, $r \leq k-3$ eintreten, die im weiteren als erfüllt vorausgesetzt wird. In diesem Falle bilden die Koeffizienten der Ableitung von $\varphi^*(s)$ die monotone Folge

$$0 < A^*_{r+1} < \ldots < (k-r-1) A^*_{k-1},$$

so daß für jede mehrfache Wurzel s^* von $\varphi^*(s)$ die Einschließung

$$0 < \min_{r+2 \leq \varkappa \leq k-1} \frac{(\varkappa-r) A}{(\varkappa-r+1) A} \leq |s^*| \leq \max_{r+2 \leq \varkappa \leq k-1} \frac{(\varkappa-r) A_{\varkappa-1}}{(\varkappa-r+1) A_\varkappa} < 1$$

gilt. Der Operator L ist also D-stabil.

§10. Weitere Sätze über absolut stabile L.

Satz 4 hat zur Folge, daß alle für D-stabile Operatoren L geltenden Sätze auch für absolut stabile gelten. Wir führen hier nur zwei an, die von besonderer Bedeutung sind. Während G.Dahlquist 1956 diese Sätze auf funktionentheoretischem Wege hergeleitet hat, können wir sie für absolut stabile Operatoren aufgrund der Darstellungen von Kapitel II elementar beweisen.

Satz 5: Für absolut stabile Operatoren L ist $p \leq k+1$.

Beweis: Es sei L absolut stabil und $p \leq k+1$. Wegen

$$\sum_{\varkappa=0}^{k-1} a_\varkappa = 1 , \quad a_\varkappa \geq 0 \quad (\varkappa = 0,\ldots,k-1)$$

liegt in

$$\alpha = \sum_{\varkappa=0}^{k-1} \alpha_{k,\varkappa} a_\varkappa = \sum_{\varkappa=0}^{k-1} \alpha_{k\varkappa} a_\varkappa$$

(vgl.(5.9), (5.10) ($\alpha\alpha$)) ein echtes Mittel der $\alpha_{k\varkappa}$ ($\varkappa = 0,\ldots,k-1$) vor. Es sind zwei Fälle zu unterscheiden:

1) $k \equiv 0\,(2)$. Nach (5.7) ist $c_{k+2} = 0$, d.h. $p > k+1$, genau dann, wenn $\alpha = 0$. Dies ist wegen (5.9),(γ) für nichtnegative Operatoren genau für $a_1 = \ldots = a_{k-1} = 0$, d.h. für den durch $\varrho(s) = -s^k + 1$ erzeugten Operator $L \in \overline{P}_{k+1}$ der Fall. Dieser ist nach Satz 4 nicht absolut stabil.

2) $k \equiv 1\,(2)$: Nach (5.10),(γ) ist $\alpha_{kk} < \alpha \leqq 0$ und damit wegen (5.7)

$$(k+2)^{-1} c_{k+2} = \sum_{\varkappa=0}^{k} \alpha_{k\varkappa} a_\varkappa = -\alpha_{kk} + \alpha > 0 ;$$

das heißt aber, es ist $p = k+1$, was zu beweisen war.

Satz 6: Für offene absolut stabile Operatoren L ist $p \leqq k$.

Beweis: Wir nehmen an, L sei absolut stabil und von einem Grade $p \geqq k+1$. Dann ist nach Satz 1 $L \in \overline{P}_{k+1}$ und nach (5.3)

$$b_k = \frac{1}{k!}\left\{ \sum_{\varkappa=0}^{k} a_\varkappa \int_0^\infty \mathfrak{N}_{k+1}(t)\,dt \right\}$$
$$= \frac{1}{k!}\left\{ \sum_{\varkappa=0}^{k-1} a_\varkappa \alpha_{k-1,\varkappa} - \alpha_{k-1,k} \right\} .$$

Hier ist

$$\alpha = \sum_{\varkappa=0}^{k-1} a_\varkappa \alpha_{k-1,\varkappa}$$

wegen

$$\sum_{\varkappa=0}^{k-1} a_\varkappa = 1 , \quad a_\varkappa \geqq 0 \quad (\varkappa = 0,\ldots,k-1)$$

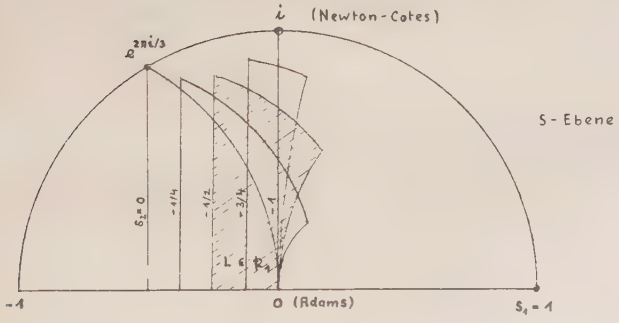
ein echtes Mittel der $\alpha_{k-1,\varkappa}$ ($\varkappa = 0,\ldots,k-1$), und nach (5.9), (5.10), (γ) folgt somit $|\alpha| < |\alpha_{k-1,k}|$, also $b_k \neq 0$. Der Operator ist demnach nicht offen, was zu zeigen war.

Anmerkungen:

1) Bislang ist jeder Versuch fehlgeschlagen, den ganzen für D-stabile Operatoren geltenden Sachverhalt (4.7) elementar herzuleiten. Ein solch elementarer Beweis ist anzustreben, weil er zeigen müßte, welche elementaren Eigenschaften der Fehlerfunktion $\mathfrak{N}_k(t)$ für die Gradbeschränkung bei stabilen Operatoren verantwortlich sind. Ist k gerade, so kann man immerhin zeigen, daß die Ungeradheit der Funktion $\mathfrak{N}_k(\frac{k}{2} + u)$ verantwortlich dafür ist, daß es D-stabile L gibt, die den Grad k+2 erreichen. Diese hat nämlich (5.9) (β) bewirkt. Ist nun $\varrho(s)$ irgendein Polynom, dessen Wurzeln sämtlich auf dem

Einheitskreis liegen und einfach sind und für das $\varrho(1) = 0$ gilt, so hat man neben (5.9),(β) mit $\varkappa = k$ noch die Beziehung $a_\varkappa = -a_{k-\varkappa}$. Für den durch $\varrho(s)$ erzeugten Operator $L \in \mathcal{L}_{k+1}$ folgt dann nach (5.7) $c_{k+2} = 0$, also $p \geq k+2$.

2) Die absolut stabilen Operatoren bilden einen kleinen Ausschnitt aus der Menge der D - stabilen. Eine Vorstellung von der Lage der charakteristischen Wurzeln bei den absolut stabilen Operatoren der Ordnung 4 vermittelt Figur 2.



Figur 2 : Lage der charakteristischen Wurzel $s_3 = \overline{s}_4$ im Falle komplexer Wurzeln bei den absolut stabilen Operatoren 4.Ordnung.

K A P I T E L IV

Optimale Integrationsverfahren.

§ 11. V o r b e m e r k u n g e n.

Mithilfe des Begriffes der absoluten Stabilität können wir nunmehr das Integrationsbeispiel von § 7 verstehen. Die Verfahren mit dem günstigen Fehlerverlauf ($0 < a_1 < 1$) sind absolut stabil, die anderen nicht. Verkleinerung der Schrittweite kann hieran nichts ändern. Man wird somit in der Praxis gut daran tun, nur absolut stabile Operatoren zu benutzen. Von diesen gibt es noch eine derart große Anzahl, daß das Problem auftritt, wie man die Menge der noch in Betracht kommenden Verfahren nach sinnvollen Argumenten weiter einschränken kann.

An dieser Stelle verdient ein Ansatz von W. Q u a d e 1957 Beachtung. Quade erzeugt eine kleine Anzahl von Operatoren L auf einem Wege, der Argumente aus der Approximationstheorie benutzt. Von diesen Operatoren könnte man meinen, daß sie aufgrund ihrer Konstruktion besondere Vorzüge besitzen müßten. Das ist jedoch nicht der Fall. Man kann zeigen, daß der Ansatz von Quade nur Verfahren L mit

$$a_{k-\varkappa} = (-1)^i a_\varkappa \quad (\varkappa = 0,1,\ldots,k)$$

liefert (wobei i der Index des Koeffizienten des trigonometrischen Polynoms ist, dessen Verschwinden gefordert wird). Nach Satz 1 und Satz 3 ist keines dieser Verfahren absolut stabil.

Im weiteren gehen wir von folgender realen Situation aus. Eine vorgelegte Differentialgleichung soll mit einer vorgegebenen Genauigkeit integriert werden. Es wird dann absolut stabile Operatoren L geben, die einen Grad $p \geq k+1$ haben und bei deren Verwendung die so gestellte Aufgabe mit einer besonders großen Schrittweite gelöst werden kann.[1)] Diese sollen dann in Bezug auf das konkrete Integrations-

---

1) vgl. hierzu auch T.E.Hull und A.C.R.Newbery 1959.

problem o p t i m a l heißen. Der näheren Bestimmung optimaler Operatoren und der größten bei ihrer Verwendung zulässigen Schrittweite dienen die weiteren Erörterungen.

Um zu glatten Ergebnissen zu kommen, ist es vorteilhaft, wenn wir uns nunmehr auf die Behandlung von Differentialgleichungen (1.1) mit

(11.1) $$f_y(x,y) = g < 0$$

beschränken (g eine Konstante). Im kleinen verhält sich $f_y(x,y)$ ohnehin bei jeder Funktion $f \in \mathfrak{F}_1$ nahezu wie eine Konstante, so daß alle für (11.1) gewinnbaren Ergebnisse für $h \to 0$ zumindest qualitative Bedeutung beibehalten. Es sei außerdem stets $k \geqslant 2$.

§ 12. Maximale Schrittweite.

Es sei $L \in \bar{\mathfrak{P}}_{k+1}$. Nach (4.5) und (5.3) ist L eindeutig durch seine Koeffizienten $a_1,\ldots,a_{k-1}$ gekennzeichnet, wenn diese einen Punkt der abgeschlossenen Hülle $\bar{\mathfrak{k}}$ des $k-1$ dimensionalen Einheitssimplex

$$\mathfrak{k} = \left\{ (a_1,\ldots,a_{k-1}) \;\middle|\; a_\varkappa > 0 \;(\varkappa = 1,\ldots,k-1);\; \sum_{\varkappa=1}^{k-1} a_\varkappa < 1 \right\}$$

beschreiben. Außerdem ist $L \in \mathfrak{P}_{k+1}$ genau dann, wenn $(a_1,\ldots,a_{k-1}) \in \mathfrak{k}$.

Es sei im weiteren $f \in \mathfrak{F}_{k+2}$; wegen (11.1) gilt dann sogar $f \in \mathfrak{J}_{k+2}$. Statt der Schrittweite h werde die Größe

(12.1) $$H = -g\,h$$

benutzt, welche die r e d u z i e r t e Schrittweite heißen soll. Ferner sollen Formel- und Rundefehler t_n bzw. r_n einer Abschätzung

(12.2) $$|t_n| + |r_n| \leqq T = T(a_1,\ldots,a_{k-1},H)$$

genügen, in der

$$T = Q_{k+2}\cdot H^{k+2} + Q_{k+3}\cdot H^{k+3}$$

ist mit in $\bar{\mathfrak{k}}$ erklärten und nichtnegativen Funktionen

$$Q_{k+2} = Q_{k+2}(a_1,\ldots,a_{k-1}),$$
$$Q_{k+3} = Q_{k+3}(a_1,\ldots,a_{k-1}),$$

für welche

$$Q_{k+2} + Q_{k+3} > 0$$

ausfällt, und die in $\bar{\mathfrak{k}}$ stetig sind.

Offenbar ist (12.2) bei festem Satz $(a_1,\dots,a_{k-1})$ eine Bedingung von der Form (9.3). Wir setzen

$$(12.3)\qquad F = F(a_1,\dots,a_{k-1},H) = T(a_1,\dots,a_{k-1},H) + K\cdot\left\{\sum_{\varkappa=0}^{k-1}|a_\varkappa+H\,b_\varkappa| - |1-H\,b_k|\right\}.$$

Hier sind a_0 und die $b_\varkappa$ vermöge (4.5) und (5.3) durch die $a_\varkappa\,(1\le\varkappa\le k-1)$ ausgedrückt zu denken.

F ist eine stetige Funktion in dem Zylinder, der durch die Bedingung

$$(a_1,\dots,a_{k-1})\in\bar{\mathfrak{k}}\ ,\quad 0\le H<\infty$$

gekennzeichnet ist, und (9.2) ist nach (8.5) eine Folge von (9.1), wenn

$$(12.4)\qquad F(a_1,\dots,a_{k-1},H)\le 0$$

gilt, und unter der Voraussetzung (12.2) nur dann.

Für $L\in\bar{\mathfrak{R}}_{k+1}$ findet man nun

$$(12.5)\qquad F(a_1,\dots,a_{k-1},0) = 0\ ,$$

$$(12.6)\qquad F(a_1,\dots,a_{k-1},H)\to+\infty \text{ für } H\to+\infty,$$

und für $L\in\mathfrak{R}_{k+1}$ ist

$$(12.7)\qquad F_H^+(a_1,\dots,a_{k-1},0) < 0\ .$$

Ferner: Setzt man

$$E_\varkappa = a_\varkappa + H\,b_\varkappa$$

und versteht man unter $H_\varkappa$ bei festem $a_\varkappa$ und $b_\varkappa\neq 0\ (0\le\varkappa\le k)$ die Nullstelle von $E_\varkappa$ bezüglich H, so gilt für $L\in\bar{\mathfrak{R}}_{k+1}$

$$(12.8)\qquad F_{HH} > 0 \quad \text{für} \quad 0<H<\infty,\ H\neq H_\varkappa,$$

während man für $H=H_\varkappa>0$, wo nur linke und rechte partielle Ableitungen von F nach H existieren, wegen

$$\frac{\partial^+|E_\varkappa|}{\partial H} - \frac{\partial^-|E_\varkappa|}{\partial H} = \begin{cases} -2\,b_\varkappa>0\ , & \text{wenn } 0\le\varkappa<k \\ 2\,b_k>0\ , & \text{wenn } \varkappa=k \end{cases}$$

die Ungleichung

$$(12.9)\qquad F_H^+(a_1,\dots,a_{k-1},H_\varkappa) - F_H^-(a_1,\dots,a_{k-1},H_\varkappa) > 0$$

herleiten kann.

Aus (12.7) bis (12.9) folgt, daß -F bei festem $(a_1,\dots,a_{k-1})\in\bar{\mathfrak{k}}$ eine konvexe Funktion in H ist und daß die Gleichung

$$(12.10)\qquad F(a_1,\dots,a_{k-1},H)=0\ ,\ (a_1,\dots,a_{k-1})\in\bar{\mathfrak{k}}$$

wegen (12.4) bis (12.6) für $(a_1,\dots,a_{k-1}) \in \mathfrak{k}$ genau eine, für $(a_1,\dots,a_{k-1}) \in \bar{\mathfrak{k}}$ höchstens eine positive Lösung

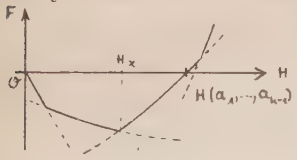
$$H = H(a_1,\dots,a_{k-1})$$

besitzt. In jedem Randpunkt von $\bar{\mathfrak{k}}$, in dem es keine solche positive Lösung gibt, werde H durch 0 definiert.

Die Funktion $H(a_1,\dots,a_{k-1})$ hat somit für $(a_1,\dots,a_{k-1}) \in \bar{\mathfrak{k}}$ die folgende grundlegende Eigenschaft, daß (vgl. auch Figur 3)

$$F(a_1,\dots,a_{k-1},H) \begin{cases} \leq 0 \text{ in } 0 \leq H \leq H(a_1,\dots,a_{k-1}) \\ > 0 \text{ für } \quad H > H(a_1,\dots,a_{k-1}) \end{cases} . \tag{12.11}$$

Aus ihr und der Stetigkeit von F kann die Stetigkeit von $H(a_1,\dots,a_{k-1})$ in $\bar{\mathfrak{k}}$ hergeleitet werden, worauf hier indes nicht eingegangen wird.



Figur 3

Die Funktion $H(a_1,\dots,a_{k-1})$ ist von großer praktischer Bedeutung. Sie gibt - grob gesagt - die obere Grenze der reduzierten Schrittweiten H an, für welche absolute Stabilität stattfindet.

Von besonderem Interesse ist dabei das Maximum von $H(a_1,\dots,a_{k-1})$ in $\bar{\mathfrak{k}}$. Jeder dieses Maximum liefernde Operator $L \in \bar{\mathfrak{K}}_{k+1}$ soll, wenn er absolut stabil ist, für das gestellte Integrationsproblem o p t i m a l heißen. Natürlich hängt die Nützlichkeit dieses Begriffes wesentlich davon ab, daß die Abschätzung (12.2) realistisch ist. Für $k = 2$ und $k = 3$ soll in den beiden nächsten Paragraphen eine eingehende Untersuchung der hier angeschnittenen Frage erfolgen.

§13. Der optimale Operator 2. Ordnung.

Die Verfahren 2. Ordnung und 3. oder höheren Grades haben die Gestalt (7.1) und sind nichtnegativ für

$$0 \leq a_1 \leq 1 , \tag{13.1}$$

was nunmehr vorausgesetzt werde. Außerdem sei $K > 0$ so klein vorausgesetzt, daß $\max_{0 \leq a_1 \leq 1} H(a_1) \leq 3/2$ ausfällt, mithin so daß von vornherein

nur das Intervall

(13.2) $\qquad 0 \leq H \leq 3/2$

betrachtet zu werden braucht. Dies bedeutet für die Praxis im allgemeinen keine besondere Einschränkung.

a) Der Formelfehler.

Es empfiehlt sich, den Formelfehler nicht wie in Paragraph 4, sondern wie folgt zu erfassen: Man entwickle Ly_n an der Stelle x_{n+1} in eine Taylorreihe unter Angabe des Restgliedes. Man findet dann

(13.3) $$t_n = \frac{a_1}{24} h^4 y^4_{n+1} + \frac{1}{72} h^5 \int_{-1}^{1} K(\tau) \cdot y^{(5)}(x_{n+1} + h\tau)\, d\tau$$

mit dem Kern

(13.4) $$K(\tau) = \begin{cases} [1 + a_1 + 3\tau]\cdot(1-\tau)^3 & \text{für } 0 \leq \tau \leq 1 \\ [1 - 2a_1 - 3(1-a_1)\tau]\cdot(1+\tau)^3 & \text{für } -1 \leq \tau \leq 0 \end{cases}$$

Dieses Vorgehen findet seine Berechtigung darin, daß $K(\tau)$ für $0 \leq a_1 \leq 1/2$ (und dies wird sich als der wichtige Bereich ergeben) nichtnegativ ist. Aus (13.3) kann man daher die Abschätzung

(13.5) $$|t_n| \leq t = \frac{a_1}{24} M_4 h^4 + \frac{2 - a_1}{180} M_5 h^5$$

für $0 \leq a_1 \leq 1/2$ gewinnen. Hier bedeuten M_4 und M_5 Schranken für $|y^{(4)}(x)|$ bzw. für $|y^{(5)}(x)|$ im Intervall $[x_n, x_{n+2}]$. Sind diese Schranken scharf, so ist (13.5) im allgemeinen auch eine scharfe Abschätzung. Im Intervall $1/2 < a_1 < 1$ gilt (13.5) im allgemeinen nicht. Darauf kommen wir später zurück. Für das erste soll diese Zeile als allgemein geltend angenommen werden.

b) Die Rundefehler.

Die Rundefehler r_n sollen in der Größenordnung von t gehalten sein. Wir setzen voraus, daß

(13.6) $$|r_n| \leq \lambda \cdot t$$

für ein festes $\lambda \geq 0$ gilt. Offenbar ist (13.6) von der Gestalt (9.3 Somit haben wir in der aus (13.5) und (13.6) folgenden Ungleichung

(13.7) $$|t_n| + |r_n| \leq T = \frac{1+\lambda}{360|g|^5} \left\{ 15|g|a_1 M_4 H^4 + (4 - 2a_1) M_5 H^5 \right\}$$

eine Abschätzung von der Gestalt (12.2) gefunden, wenn $M_4 + M_5 > 0$ vorausgesetzt wird.

c) Die Funktion $H(a_1)$.

Wie man (7.1) entnimmt, gilt unter Benutzung der Bezeichnungen von Paragraph 12

$$E_0 = 1 - a_1 + \frac{4 - 5a_1}{12} \cdot H ,$$

$$E_1 = a_1 + \frac{16 - 8a_1}{12} \cdot H ,$$

$$E_2 = -1 + \frac{4 + a_1}{12} \cdot H .$$

Wir definieren folgende Punktmengen:

$$\mathfrak{X}_1 = \left\{ (a_1, H) \;\middle|\; 0 \le a_1 \le 1 ;\; 0 \le H \le \frac{3a_1}{4 - 2a_1} \right\} ,$$

$$\mathfrak{X}_2 = \left\{ (a_1, H) \;\middle|\; 0 \le a_1 \le 1 ;\; \frac{3a_1}{4 - 2a_1} \le H \le 3/2 \right\} .$$

Offenbar ist $\mathfrak{X}_1 \cup \mathfrak{X}_2$ das ganze durch (13.1),(13.2) beschriebene Rechteck $\mathfrak{R}$. Der gemeinsame Rand von $\mathfrak{X}_1$ und $\mathfrak{X}_2$ ist die Kurve

(13.8) $\mathfrak{E}$: $\qquad H = \frac{3a_1}{4 - 2a_1} , \quad 0 \le a_1 \le 1 ,$

deren Verlauf aus Figur 4 hervorgeht.

In $\mathfrak{R}$ ist $E_2 < 0$, also $|E_2| = -E$. Ferner erhält man nach einiger Rechnung

$$|E_0| + |E_1| = \begin{cases} 1 + \frac{20 - 13a_1}{12} \cdot H & \text{in } \mathfrak{X}_1 \\ 1 - 2a_1 - \frac{4 - a_1}{4} \cdot H & \text{in } \mathfrak{X}_2 \end{cases}$$

Man findet daher aus (12.3) unter Berücksichtigung von (13.7) für die Funktion

$$\phi = \phi(a_1, H) = 360 |g|^5 \cdot F(a_1, H)$$

die Darstellung

(13.9) $$\phi = (1+\lambda) \cdot \left[(4 - 2a_1) M_5 H^5 + 15 |g| a_1 M_4 H^4 \right] + \begin{cases} 360 (2 - a_1) |g|^5 K H & \text{in } \mathfrak{X}_1 \\ -120 (2 - a_1) |g|^5 K H - 720 a_1 g^5 K & \text{in } \mathfrak{X}_2 \end{cases} ,$$

und $H = H(a_1)$ genügt der mit (12.10) gleichwertigen Gleichung

$$\phi(a_1, H) = 0, \qquad 0 \le a_1 \le 1 .$$

Man kann zeigen: In jedem Intervall, in welchem $(a_1, H(a_1)) \in \mathfrak{X}_1$ ($\in \mathfrak{X}_2$) gilt, ist $H(a_1)$ eigentlich monoton fallend (wachsend). $H(a_1)$ nimmt demzufolge sein Maximum $\bar{H}$ in $[0,1]$ in genau einem (inneren) Punkte u an, der durch die Nebenbedingung $(u, H(u)) \in \mathfrak{E}$ gekennzeichnet ist. Unser Ziel ist nun, u und $\bar{H} = H(u)$ genauer zu bestimmen.

Wir setzen

$$D_2 = \frac{(1+\lambda)\,(M_5 + 5|g|M_4)}{|g|^5 \cdot K}$$

und nennen diese Größe die Genauigkeit der Rechnung. Es gilt dann der

Satz 7: Ist $D_2 \geq 2880$,[1] so ist die größte reduzierte Schrittweite, für welche mit einem absolut stabilen Operator L der 2. Ordnung und vom (höchsten) Grad $p = 3$ die Genauigkeit D_2 erzielt wird, gegeben durch

$$\bar{H} = \sqrt[4]{\frac{180}{D_2}}\,.$$

Der diese reduzierte Schrittweite ermöglichende einzige Operator ist gekennzeichnet durch $a_1 = u$ mit

$$u = \frac{4\bar{H}}{3 + 2\bar{H}}\,.$$

Beweis: Bestimmt man $\bar{H}$ und u gemäß Satz 7, so ist $0 < H \leq 1/2$ und $0 < u \leq 1/2$, d.h. der durch $a_1 = u$ bestimmte Operator $L \in \bar{\mathfrak{P}}_3$ ist positiv und somit nach Satz 2 und Satz 5 absolut stabil und vom genauen Grad $p = 3$. Das Wertepaar ist gemeinsame Lösung von (13.8) und (13.9); nach dem oben Gesagten ist also

$$\bar{H} = H(u) = \max_{0 \leq a_1 \leq 1} H(a_1)\,,$$

und dieses Maximum wird nur von diesem Operator $\in \bar{\mathfrak{P}}_{k+1}$ erreicht. Zugleich gilt für L - wie für alle Operatoren aus $\bar{\mathfrak{P}}_{k+1}$ mit $0 \leq a_1 \leq 1/2$ - die Abschätzung (13.5). Für $1/2 < a_1 \leq 1$ jedoch muß die rechte Seite von (13.5) durch irgendeinen größeren Wert ersetzt werden. Man findet dann aufgrund von (12.3) und (12.11), daß die Funktion $H(a_1)$, wie sie sich unter Zugrundelegen von (13.5) ergab, für $1/2 < a_1 \leq 1$ eine reduzierte Schrittweite angibt, die bereits zu groß ist, um die Genauigkeit D_2 zu verbürgen. Damit ist gezeigt, daß die Genauigkeit D_2 mit der reduzierten Schrittweite $\bar{H}$ nur mit dem durch $a_1 = u$ gekennzeichneten Operator erzielt werden kann.

---

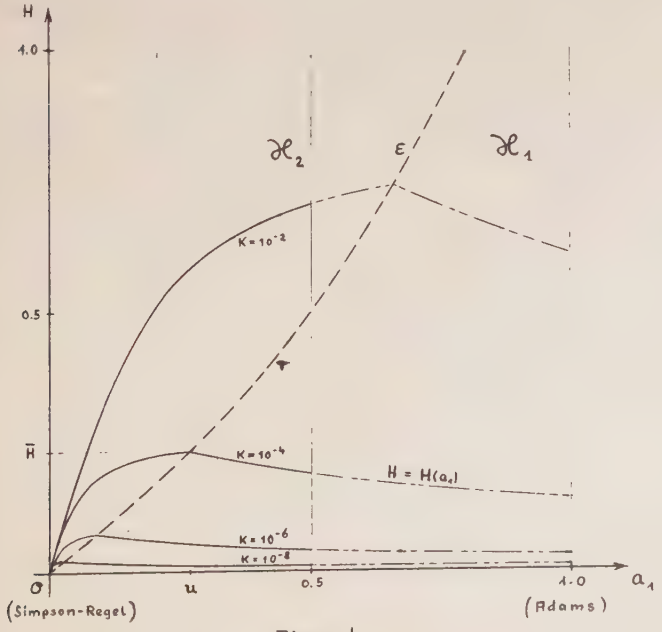
1) In der Praxis werden im allgemeinen Genauigkeiten der Größenordnung 10^{-5} oder 10^{-10} in Betracht kommen, so daß diese Voraussetzung keine besondere Einschränkung bedeutet.

Der genaue Verlauf von $H(a_1)$ ist für den Fall $M_4 = M_5 = -g = 1$ und für verschiedene Werte von K in Figur 4 dargestellt. Dieser Fall ist, wie an weiteren Beispielen bemerkt wurde, durchaus charakteristisch.

Anmerkungen :

1) Das Integrationsbeispiel von Paragraph 7 findet nun durch Satz 7 auch eine quantitative Erklärung. Nimmt man $M_4 = M_5 = 1$, so wird mit $\lambda = 0$ die dadurch gegebene Genauigkeit $6/K$ für $a_1 = u = 0.27$ erzielt, und zwar mit der maximal zulässigen Schrittweite $\bar{h} = 0.23$. Wie das Beispiel zeigt, liefert der so nach Satz 7 bestimmte Operator den bei weitem günstigsten Fehlerverlauf.

2) Verlangt man hohe Genauigkeit, so fällt $\bar{H}$ naturgemäß klein aus. Gleichzeitig aber wird auch u klein. Für $D_2 \to \infty$ gilt $u \to 0$, das heißt, der optimale Operator liefert schließlich die Simpson - Regel. Entsprechendes gilt insbesondere für $g \to 0$, wenn $M_4 + M_5 \geq M > 0$ und K fest gehalten werden, so daß $D_2 \to \infty$. Die Simpson - Regel nimmt



Figur 4

somit - selbst nicht absolut stabil - eine besondere Grenzstellung bezüglich der absolut stabilen Operatoren ein. Mit einigem Recht kann man sie "optimal für $h \to 0$" nennen.

3) Bei der Differentialgleichung $y' = g\,y$, $g < 0$, gilt entlang eines jeden Integrals

$$M_4 = g^4 |y|, \quad M_5 = |g|^5 |y|$$

und man erhält für $y(x) \neq 0$

$$\bar{H} = \sqrt[4]{\frac{30}{1+\lambda}} \cdot \frac{K'}{|y|} .$$

$\bar{H}$ hängt in diesem Falle also nur vom maximal zulässigen relativen $K/|y|$ ab.

4) An der Art, wie λ in D_2 und damit in $\bar{H}$ eingeht, erkennt man, daß es keinen Sinn hat, λ sehr klein zu wählen, vielmehr, daß eher sogar eine gewisse Großzügigkeit bezüglich der Rundefehler angebracht ist.

§14. Der optimale Operator 3. Ordnung.

Im Falle $k = 3$ sind die nichtnegativen Operatoren vierten und höheren Grades gekennzeichnet durch

$$\begin{array}{ll} a_0 = 1 - a_1 - a_2 & b_0 = (9\,a_1 + 8\,a_2 - 9)/24 \\ a_1 : \text{Parameter} & b_1 = (19\,a_1 + 32\,a_2 - 27)/24 \\ a_2 : \text{Parameter} & b_2 = (-5\,a_1 + 8\,a_2 - 27)/24 \\ a_3 = -1 & b_3 = (a_1 - 9)/24 \end{array} \tag{14.1}$$

wobei (a_1, a_2) der abgeschlossenen Hülle $\bar{\ell}$ des Dreiecks

$$\ell = \{(a_1, a_2) \mid a_1 > 0,\ a_2 > 0,\ a_1 + a_2 < 1\} \tag{14.2}$$

angehören muß. (14.1) erhält man dabei aus (5.3).

Es werde vorausgesetzt, daß $K > 0$ so klein sei, daß

$$\max_{\bar{\ell}} H(a_1, a_2) < 1$$

ausfällt, so daß von vornherein nur das Intervall

$$0 \leq H \leq 1 \tag{14.3}$$

in Betracht gezogen zu werden braucht.

a) Der Formelfehler.

Der Formelfehler werde mithilfe von (4.3) und (4.4), genommen für $q = 3$, abgeschätzt.

D er Kern $K_3(\tau)$ ist zweimal stetig differenzierbar, und es gilt

$$K_3(i) = K_3'(i) = K_3''(i) = 0 \quad \text{für} \quad i = 0, 3. \tag{14.4}$$

Für $(a_1, a_2) \in \bar{\mathfrak{E}}$ - und nur solche Wertepaare werden weiterhin in Betracht gezogen - gilt

$$K_3(2) > 0, \quad K_3(1) = 3(1 - a_1 - a_2) + a_2 \geqq 0,$$

wobei im letzten Falle das Gleichheitszeichen nur für $a_1 = 1$, $a_2 = 0$ gesetzt werden darf. Wir zeigen, daß überhaupt

$$K_3(\tau) \geqq 0 \quad \text{in } [0,3]$$

gilt.

Nach (4.4) ist $K_3(\tau)$ in jedem Intervalle $j-1 \leqq \tau \leqq j$ $(j = 1,2,3)$ durch ein Polynom $P_j(\tau)$ vierten Grades erklärt mit dem bei τ^4 stehenden Koeffizienten $\sum_{i \geq j}^{\infty} a_i$. Wir haben zu beweisen, daß für jedes j $(j = 1,2,3)$ $P_j(\tau) \geqq 0$ in $[j-1, j]$.

Zunächst sei $j = 1$. Entweder es ist $a_1 + a_2 - 1 = 0$. Dann entartet $P_1(\tau)$ in ein Polynom höchstens dritten Grades, welches wegen (14.4) die dreifache Nullstelle $\tau = 0$, in $[0,1]$ also keinen Zeichenwechsel hat. Oder es ist $a_1 + a_2 - 1 < 0$. Dann liegt die vierte Wurzel von $P_1(\tau)$ wegen $P_1(1) > 0$ und $P_1(\tau) \to -\infty$ für $\tau \to \infty$ in dem Intervall $(1, \infty)$. In beiden Fällen hat also $P_1(\tau)$ in $[0,1]$ keinen Zeichenwechsel. Aus $a_1 = 1$, $a_2 = 0$ folgt $P_1(1) = 0$ und nach dem oben Gesagten $P_1(\tau) \equiv 0$. In jedem anderen Falle ist $P_1(1) > 0$ und damit $P_1(\tau) \geqq 0$ in $[0,1]$, womit allgemein $P_1(\tau) \geqq 0$ in $[0,1]$ gilt.

Den Fall $j = 3$ kann man ganz entsprechend behandeln.

Es bleibt der Fall $j = 2$ zu untersuchen. Man kann nachrechnen, daß

$$\frac{2-\tau}{4} P_2'(\tau) + P_2(\tau) = -b_2(2-\tau)^3 - (3-\tau)^3[1 + b_3(6-\tau)].$$

An (14.1) erkennt man, daß

$$b_2 < 0, \quad b_3 \leqq -1/3.$$

Somit erhält man für $1 \leqq \tau \leqq 2$ die Ungleichung

$$\frac{2-\tau}{4} P_2'(\tau) + P_2(\tau) \geqq (3-\tau)^3/3 > 0.$$

In [1,2] folgt somit aus $P_2'(\tau) = 0$, daß $P_2(\tau) \geqq 0$ ist. Wegen $P_2(1) \geqq 0$ und $P_2(2) > 0$ gilt dann aber $P_2(\tau) \geqq 0$ in dem ganzen Intervalle [1,2], was noch zu zeigen war.

Wir wissen nunmehr, daß $K_3(\tau)$ in [0,3] nichtnegativ ist. Berücksichtigt man, daß

$$\int_0^3 K_3(\tau)\, d\tau = c_5 > 0$$

gilt - was man anhand des Falles $y^{(5)}(x) = \text{const}$ leicht erkennt -, so erhält man aus (4.3) für alle Operatoren $L \in \bar{\mathcal{P}}_4$ zuletzt die Abschätzung des Formelfehlers

$$|t_n| \leqq t = \frac{1}{5!}\, c_5\, M_5\, h^5 . \tag{14.5}$$

b) Die Rundefehler.

Wie in Paragraph 13 sollen die Rundefehler einer Abschätzung

$$|r_n| \leqq \lambda \cdot t \tag{14.6}$$

mit einer nichtnegativen Konstanten λ genügen. Man hat dann

$$|t_n| + |r_n| \leqq T = \frac{1}{5!}\, K \cdot D_3 \cdot c_5 \cdot H^5 , \tag{14.7}$$

wobei die Größe

$$D_3 = \frac{(1+\lambda)\, M_5}{K\, |g|^5} \tag{14.8}$$

wieder die Genauigkeit der Rechnung heißen soll. Ist $M_5 > 0$, was vorausgesetzt wird, so hat (14.7) die Gestalt (12.2).

c) Die Funktion $H(a_1, a_2)$.

Es sei

$$E_i^* = E_i^*(a_1, a_2, H) = \frac{24}{K} \cdot E_i \quad (i = 0,1,2,3),$$
$$F^* = F^*(a_1, a_2, H) = \frac{24}{K} \cdot F .$$

Aus (12.3) erhält man bei Berücksichtigung von (14.7), (14.1) und (5.7) wegen

$$(\alpha_{30}, \alpha_{31}, \alpha_{32}, \alpha_{33}) = (0, -\tfrac{19}{30}, -\tfrac{4}{15}, -\tfrac{9}{10}) ,$$

daß

$$F^* = \frac{D_3}{30}\,(27 - 19\,a_1 - 8\,a_2)\,H^5 + E_3^* + \sum_{i=0}^{2} |E_i^*|$$

gilt mit

$$E_0^* = 24(1 - a_1 - a_2) + H\cdot(9\,a_1 + 8\,a_2 - 9) ,$$
$$E_1^* = 24\,a_1 + H\cdot(19\,a_1 + 32\,a_2 - 27) ,$$

$$E_2^* = 24\,a_2 + H\cdot(-5\,a_1 + 8\,a_2 - 27),$$
$$E_3^* = -24 + H\cdot(a_1 - 9).$$

Die $H(a_1,a_2)$ definierende Gleichung (12.10) geht über in die gleichwertige Gleichung

(14.9) $$F^*(a_1,a_2,H) = 0, \quad (a_1,a_2) \in \bar{\mathfrak{k}}.$$

Wir beweisen jetzt den

Satz 8: Ist $D_3 \geq 52$, so ist die größte reduzierte Schrittweite $\bar{H}$, bei der mit einem absolut stabilen Operator dritter Ordnung und vom (höchsten) Grad $p = 4$ die Genauigkeit D_3 erzielt werden kann, gekennzeichnet als die einzige im Intervall $(0,1)$ gelegene Wurzel der algebraische Gleichung

$$\frac{D_3}{60}(3H^7 + H^6 - 28H^5 - 12H^4) + H^3 - 3H^2 - 6H + 36 = 0.$$

Der einzige diese reduzierte Schrittweite $\bar{H}$ ermöglichende absolut stabile Operator ist gegeben durch $(a_1,a_2) = (u,v)$,

$$u = \frac{36 - 42\bar{H} + 9\bar{H}}{36 + 6\bar{H} - 7\bar{H}},$$
$$v = \frac{(27 + 5u)\cdot\bar{H}}{24 + 6\bar{H}}.$$

Für $D_3 \to \infty$ gilt

$$H \sim \sqrt[4]{\frac{180}{D_3}}.$$

Beweis: Wir nehmen an, es sei $D_3 \geq 52$, und haben zu zeigen, daß $H(a_1,a_2)$ sein Maximum bezüglich $\bar{\mathfrak{k}}$ in genau einem inneren Punkte von $\bar{\mathfrak{k}}$ annimmt. Danach müssen dieses Maximum und jener innere Punkt genau bestimmt werden. Außerdem ist zu zeigen, daß die zunächst zu treffende Voraussetzung $\bar{H} < 1$ von selbst erfüllt ist.
Der Beweis ist recht mühsam und wird sich bis S.47 erstrecken. Verschiedene Einzelheiten, die genau überlegt werden können, werden nur mit einem kurzen Hinweis berücksichtigt.

Ist $(x,y) = P \in \bar{\mathfrak{k}}$, so setzen wir

$$\mathfrak{y} = \mathfrak{y}(P) = \{(a_1,a_2) \mid 19a_1 + 8a_2 = 19x + 8y\}$$

und

$$\omega = \omega(P) = \mathfrak{y} \cap \bar{\mathfrak{k}}.$$

$\omega(P)$ ist im allgemeinen eine Strecke, kann aber auch in einen Punkt entarten. Wir beweisen jetzt zwei Hilfssätze.

Hilfssatz 1: Wenn für ein Wertetripel (x,y,Z) mit

$$P = (x,y) \in \overline{\xi} \quad , \quad 0 \leqq Z = H(x,y) < 1$$

$E_2^*(x,y,Z) = 0$ ausfällt, so gilt

$$\max_{\omega(P)} H(a_1,a_2) = Z \, .$$

$H(a_1,a_2)$ nimmt den Wert Z auf ω nur in P an, und nur in diesem Punkt von ω gilt $E_2^*(a_1,a_2,H(a_1,a_2)) = 0$.

Beweis: Die Voraussetzungen des Hilfssatzes 1 seien erfüllt für x,y,Z. Ist $\omega = \{P\}$, so ist nichts zu beweisen. Im weiteren habe ω die Darstellung

$$a_2 = a_2(a_1) \, , \quad \xi \leqq a_1 \leqq \eta \text{ mit } \xi < \eta \, .$$

Setzt man für $\xi \leqq a_1 \leqq \eta$

$$\varrho = \varrho(a_1) = E_2^*(a_1,a_2(a_1),Z) \, ,$$

$$\varphi = \varphi(a_1) = F^*(a_1,a_2(a_1),Z) \, ,$$

so kann man nachprüfen, daß in $(\xi,\eta]$ bzw. in $[\xi,\eta)$ die linken bzw. rechten Ableitungen von φ gebildet werden können und daß deren Zeichen durch das der entsprechenden Ableitung von $|\varrho|$ bestimmt wird. Man findet so

$$\varphi'_+(a_1) > 0 \, , \text{ wenn } x \leqq a_1 \leqq \eta \, ,$$

$$\varphi'_-(a_1) < 0 \, , \text{ wenn } \xi \leqq a_1 \leqq x \, .$$

Wegen $y = a_2(x)$ ist $\varphi(x) = 0$ und gilt

$$F^*(a_1,a_2(a_1),Z) = \varphi(a_1) > 0 \quad \text{für } x \neq a_1 \in [\xi,\eta] \, .$$

Wegen (12.11) folgt daraus

$$H(a_1,a_2(a_1)) < Z \quad \text{für} \quad x \neq a_1 \in [\xi,\eta] \, ,$$

was zu beweisen war. Die Annahme, daß in zwei verschiedenen Punkten von ω $E_2^*(a_1,a_2,H(a_1,a_2))$ verschwindet, führt sofort zum Widerspruch zum bisher Bewiesenen.

Hilfssatz 2: Wenn für ein Wertetripel (x,y,Z) mit

$$P = (x,y) \in \overline{\xi} \quad , \quad 0 \leqq Z = H(x,y) < 1$$

$\omega(P)$ eine Darstellung

$$a_2 = a_2(a_1) \quad , \quad x \leqq a_1 \leqq \eta$$

hat und wenn

$$E_2^*(a_1,a_2,H(a_1,a_2)) \leqq 0 \quad \text{auf } \omega(P),$$

so gilt

$$\max_{\omega(P)} H(a_1,a_2) = Z.$$

Beweis: Die Voraussetzungen des Hilfssatzes 2 seien für x,y,Z erfüllt. Ist $\omega = \{P\}$, so ist nichts zu beweisen. Im anderen Falle setze man

$$\ell = \ell(a_1) = E_2^*(a_1,a_2(a_1),Z),$$
$$\varphi = \varphi(a_1) = F^*(a_1,a_2(a_1),Z).$$

Das Zeichen von φ'_+ wird durch das von $\frac{\partial^+}{\partial a_1}|\ell|$ bestimmt. Wegen $\ell \leqq 0$ findet man

$$\varphi'_+ > 0 \quad \text{für } a_1 \in [x,\eta);$$

und weil $y = a_2(x)$ ist und $\varphi(x) = 0$, so kann man folgern, daß

$$F^*(a_1,a_2(a_1),Z) = \varphi(a_1) > 0 \text{ für } x \neq a_1 \in [x,\eta].$$

Wegen (12.11) ergibt sich somit

$$H(a_1,a_2(a_1)) < Z \quad \text{für } x \neq a_1 \in [x,\eta],$$

was zu beweisen war.

Nach diesen Vorbereitungen wenden wir uns dem eigentlichen Teile des Beweises von Satz 8 zu. Es sei Ω die Menge aller Punkte aus $\bar{\mathfrak{E}}$, in denen $E_2^*(a_1,a_2,H(a_1,a_2)) = 0$ gilt.

1) Die Bedingung $E_2^*(a_1,a_2,H) = 0$ ist gleichwertig mit

(14.10) $$a_2 = \chi(a_1,H),$$

wenn

$$\chi(a_1,H) = \frac{(27+5a_1)\cdot H}{24+8H}$$

gesetzt wird. In dem von uns betrachteten Quadrat:

$$0 \leqq a_1 \leqq 1, \quad 0 \leqq H \leqq 1$$

ist $\chi(a_1,H)$ in Bezug auf H eigentlich monoton wachsend. Es gilt ferner mit $0 \leqq H = H(a_1,a_2) \leqq 1$ genau dann $(a_1,\chi(a_1,H)) \in \bar{\mathfrak{E}}$, wenn

(14.11) $$(a_1,H) \in \mathfrak{M} = \left\{(a_1,H) \mid 0 \leqq a_1 \leqq \frac{24-19H}{24+15H},\ 0 \leqq H \leqq 1\right\}$$

gilt.

Man kann daher Ω kennzeichnen als die Menge der Punkte (a_1,a_2), für die $H(a_1,a_2)$ die Bedingungen (14.10) und (14.11) erfüllt.

In Ω ist (14.9) wegen (14.10) gleichwertig mit der Gleichung

(14.12) $\qquad \phi = \phi(a_1, H) = 0 \quad , \qquad (a_1, H) \in \mathfrak{M},$

wobei

$$\phi = (3+H)\cdot F^*(a_1, \chi(a_1, H), H)$$

zu setzen ist.

Wir bezeichnen mit ε_j $(j = 0,1)$ die beiden Kurven

$$\varepsilon_0 : \quad a_1 = \varphi_0(H) = \frac{36 - 42H + 9H^2}{36 + 6H - 7H^2}$$

$$\varepsilon_1 : \quad a_1 = \varphi_1(H) = \frac{27\cdot H\cdot(1-H)}{24 + 27H + 13H^2} \qquad (0 < H < 1);$$

auf die sehr einfache Diskussion des Verlaufs der beiden Kurven verzichten wir und verweisen stattdessen auf Figur 5.

Die beiden Kurven ε_0 und ε_1 zerlegen $\mathfrak{M}$ in drei Teilgebiete, deren abgeschlossene Hüllen mit $\mathfrak{K}_0$, $\mathfrak{K}_1$, $\mathfrak{K}_2$ bezeichnet werden sollen. Die Numerierung sei dabei so gewählt, daß für die Funktionen

$$\mathcal{E}_j = \mathcal{E}_j(a_1, H) = E_j^*(a_1, \chi(a_1, H), H) \qquad (j = 0,1)$$

die folgenden Relationen erfüllt sind:

$$\mathcal{E}_0 \leqq 0 \ , \quad \mathcal{E}_1 \geqq 0 \quad \text{auf } \mathfrak{K}_0,$$

$$\mathcal{E}_0 \geqq 0 \ , \quad \mathcal{E}_1 \leqq 0 \quad \text{auf } \mathfrak{K}_1,$$

$$\mathcal{E}_0 \geqq 0 \ , \quad \mathcal{E}_1 \geqq 0 \quad \text{auf } \mathfrak{K}_2$$

(vgl. Figur 5). Nunmehr kann ϕ explizit angegeben werden. Es ist

$$\phi = \phi_i = \phi_i(a_1, H) \quad \text{für} \quad (a_1, H) \in \mathfrak{K}_i \ (i = 0,1)$$

mit

$$\phi_0 = \frac{D_3}{30} H^5 \cdot \left\{ 81 - a_1(57+24H) \right\} - 144 - 48H + 54H^2 + a_1(144 + 96H + 26H^2)$$

$$\phi_1 = \frac{D_3}{30} H^5 \cdot \left\{ 81 - a_1(57+24H) \right\} - 54H - 72H^2 - a_1(144 + 90H + 24H^2),$$

$$\phi_2 = \frac{D_3}{30} H^5 \cdot \left\{ 81 - a_1(57+24H) \right\} - 216H + 90H^2 + a_1(72H + 54H^2).$$

2) Es werde jetzt die Gleichung

(14.12)a $\qquad \phi(a_1, \psi) = 0 \ , \quad 0 \leqq a_1 \leqq 1, \quad 0 \leqq \psi \leqq 1$

betrachtet. Wegen

$$\phi(a_1, 0) = 0 \ ,$$

$$\phi_H^+(a_1, 0) < 0 \ , \qquad (0 < a_1 < 1)$$

$$\phi(a_1, 1) > 0$$

und der Stetigkeit von ϕ gibt es in $0 < a_1 < 1$ für (14.12)a eine kleinste positive Lösung $\psi(a_1)$.

Setzt man

(14.13) $\chi(a_1) = \chi(a_1, \psi(a_1))$,

so gilt, wenn $(a_1, \psi(a_1)) \in \mathfrak{M}$, also $(a_1, \chi(a_1)) \in \bar{\mathfrak{E}}$,

$F^*(a_1, \chi(a_1), \psi(a_1)) = 0$,

und wegen (12.11) folgt hieraus

(14.14) $H(a_1, \chi(a_1)) = \psi(a_1)$.

Nach Konstruktion ist dann also

(14.15) $(a_1, \chi(a_1)) \in \Omega$, wenn $(a_1, \psi(a_1)) \in \mathfrak{M}$.

Die Umkehrung gilt unter Umständen nicht. Aus der Monotonie von $\chi(a_1, H)$ in H und der Minimaleigenschaft von $\psi(a_1)$ folgt jedoch unter der Bedingung $0 < a_1 < 1$

(14.16) $a_2 \geq \chi(a_1)$, wenn $(a_1, a_2) \in \Omega$, $a_2 > 0$.

3) Nunmehr werde die Funktion $\psi(a_1)$ genauer untersucht. Man kann zeigen, daß in dem durch

$0 < a_1 < 1$, $0 < \psi < 1$

gekennzeichneten Gebiet die Beziehungen

$H \cdot \phi_{i_H} - \phi_i > 0$ für $i = 0,2$,

$H \cdot \phi_{1_H} - 2\phi_1 > 0$,

$\{81 - a_1(57+24H)\} \cdot (a_1 \phi_{i\,a_1} - \phi_i) + 81\,\phi_i < 0$ für $i = 1,2$,

$(1 - a_1) \cdot (a_1 \phi_{o\,a_1} - \phi_o) + \phi_o > 0$

gelten. Anwendung des Satzes über implizite Funktionen ergibt bei Berücksichtigung der Minimaleigenschaft von $\psi(a_1)$ aus (14.12)a, daß $\psi(a_1)$ in (0,1) stetig ist und daß es genau ein $u \in (0,1)$ gibt, so daß

$\psi(a_1)$ { in $(0,u]$ eigentlich monoton wächst, in $[u,1)$ eigentlich monoton fällt.

Hieraus geht unter anderem hervor, daß $\psi(a_1)$ stetig in den Rand von (0,1) fortgesetzt werden kann, was nunmehr geschehe. (14.14) bis (14.16) gelten dann wegen der Stetigkeit der vorkommenden Funktionen auch noch für $a_1 = 0$ und $a_1 = 1$. Jedoch gilt für diese Werte nur noch $\psi(a_1) \geq 0$.

$\psi(a_1)$ hat in $[0,1]$ ein Maximum Ψ mit $0 < \Psi < 1$. Dieses wird nach dem oben Gesagten nur für $a_1 = u$ angenommen, wobei das Wertepaar (u, Ψ)

durch die Beziehung $(u,\psi) \in \mathcal{E}_0$ oder, was dasselbe ist, durch

(14.17) $\qquad u = \varphi_0(\psi)$

gekennzeichnet ist. Man kann zeigen, daß $\mathcal{E}_0$ nur innere Punkte von $\mathfrak{M}$ enthält, so daß insbesondere (u,ψ) ein innerer Punkt von $\mathfrak{M}$ ist. Ferner gilt $(a_1,\psi(a_1)) \in \mathfrak{M}$ für $0 \leqq a_1 \leqq u$. Es sei $[0,\alpha]$ das größte in $[0,1]$ enthaltene Intervall dieser Art. Sicher ist dann $\alpha > u$. Außerdem gilt $(\alpha,\psi(\alpha)) \in \mathcal{E}$, wenn mit $\mathcal{E}$ der Teil des Randes von $\mathfrak{M}$ bezeichnet wird, in dem

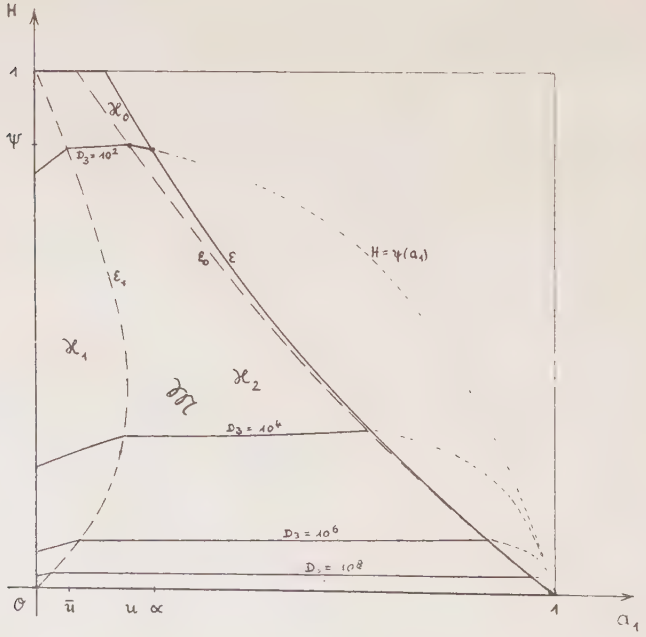
$$a_1 = \frac{24 - 19\,H}{24 + 13\,H}, \quad 0 \leqq H \leqq 1$$

gilt.

Es ist möglich, daß $\alpha < 1$ ist. In diesem Fall kann gezeigt werden, daß $(a_1,\psi(a_1)) \notin \mathcal{E}$ für $\alpha < a_1 < 1$ gilt. Wegen der Maximaleigenschaft von $[0,\alpha]$ und der Stetigkeit von $\psi(a_1)$ folgt daraus dann

$$(a_1,\psi(a_1)) \notin \mathfrak{M} \quad \text{für} \quad \alpha < a_1 < 1.$$

Man vergleiche hierzu Figur 5.



Figur 5

4) Man betrachte schließlich die Punktmenge

$$\Omega^* = \{(a_1,a_2) \mid a_2 = \chi(a_1),\ 0 \leqq a_1 \leqq \alpha\}.$$

Wegen (14.15) ist diese Menge in Ω enthalten. Es seien

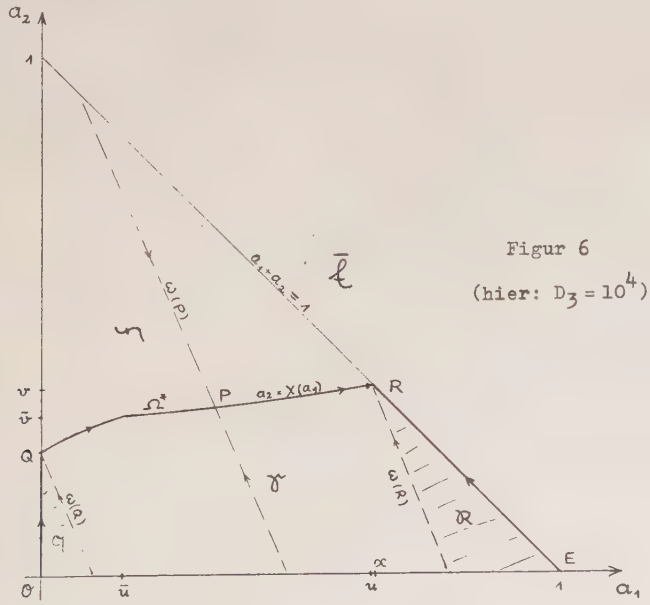
$$\mathcal{O} = (0,0), \qquad E = (1,0),$$
$$Q = (0,\chi(0)), \qquad R = (\alpha,\chi(\alpha)),$$

und mit $\mathcal{Q}$ bzw. $\mathcal{R}$ sollen die kompakten Dreiecke mit den Ecken $\mathcal{O}$ bzw. E und den Seiten $\omega(Q)$ bzw. $\omega(R)$ bezeichnet werden.[1] Wegen $(\alpha,\psi(\alpha)) \in \mathcal{E}$ liegt R auf der Geraden $a_1 + a_2 = 1$, und da Ω^* offenbar jede "zwischen" $\omega(Q)$ und $\omega(R)$ gelegene Strecke ω trifft, nach Hilfssatz 1 jedoch keine von ihnen in mehr als einem Punkt, so gilt, wenn

$$\mathcal{F} = \bigcup_{P \in \Omega^*} \omega(P)$$

gesetzt wird,

(14.18) $$\bar{\mathcal{E}} = \mathcal{F} \cup \mathcal{Q} \cup \mathcal{R}$$ (vgl. Figur 6).



Figur 6

(hier: $D_3 = 10^4$)

1) $\mathcal{Q}$ und $\mathcal{R}$ können auch in die aus dem einen Punkt $\mathcal{O}$ bzw. E bestehenden Punktmengen entarten.

Ferner enthält das durch Ω^* und den Polygonzug $Q_0 E R$ berandete Gebiet $\mathfrak{G}$ wegen (14.16) und des am Ende von 3) Gesagten keinen Punkt von $\mathfrak{L}$. Man kann zeigen, daß $E_2^*(\frac{4}{3}, a_2, H(\frac{4}{3}, a_2)) < 0$ ist für alle hinreichend kleinen $a_2 > 0$. Aus beidem zusammen und der Stetigkeit von E_2^* folgt, daß in der abgeschlossenen Hülle von $\mathfrak{G}$ und damit in $\mathfrak{H} \cup \mathfrak{R}$

$$E_2^*(a_1, a_2, H(a_1, a_2)) \leq 0$$

gilt. Man kann daraus folgern, daß $H(a_1, a_2)$ auf $\overline{\mathfrak{G}\mathfrak{Q}}$ bzw. auf $\overline{E R}$ die Gleichung

$$0 = F_1 = \frac{D_3}{30}(27 - 8a_2)H^5 + (36 - 32a_2)H - 48a_2$$

bzw.

$$0 = F_2 = \frac{D_3}{30}(19 - 11a_1)H^5 + 16H - 48(1 - a_1)$$

erfüllt. Da für $0 < H < 1$

$$H \cdot F_{1H} - F_1 > 0, \qquad F_{1a_2} < 0,$$
$$H \cdot F_{2H} - F_2 > 0, \quad (1-a_1)(a_1 \cdot F_{2a_1} - F_2) + F_2 > 0$$

gilt, kann gefolgert werden, daß $H(a_1, a_2)$ auf $\overline{\mathfrak{G}\mathfrak{Q}}$ mit a_2 monoton wächst und auf $\overline{E R}$ in Abhängigkeit von a_1 monoton fällt. Mithin erhält man bei Berücksichtigung des unter 3) Gesagten und von (14.14)

$$\max_{\overline{\mathfrak{G}\mathfrak{Q}}} H(a_1, a_2) = H(0, X(0)) = \psi(0) < \Psi,$$
$$\max_{\overline{E R}} H(a_1, a_2) = H(\alpha, X(\alpha)) = \psi(\alpha) < \Psi.$$

Andererseits folgt aus Hilfssatz 2

$$\max_{\mathfrak{H}} H(a_1, a_2) = \max_{\overline{\mathfrak{G}\mathfrak{Q}}} H(a_1, a_2),$$
$$\max_{\mathfrak{R}} H(a_1, a_2) = \max_{\overline{E R}} H(a_1, a_2)$$

und nach Hilfssatz 1 und (14.14) ist

$$\max_{\mathfrak{F}} H(a_1, a_2) = \max_{\Omega^*} H(a_1, a_2) = \max_{0 \leq a_1 \leq \alpha} \psi(a_1) = \Psi.$$

Wegen (14.18) erhält man so insgesamt

$$\overline{H} = \max_{\mathfrak{E}} H(a_1, a_2) = \Psi,$$

und dieses Maximum wird nur auf Ω^* und nur für $a_1 = u$ angenommen. Wegen (14.17) genügt $\overline{H}$ der Gleichung $\Phi(\varphi_0(\overline{H}), \overline{H}) = 0$, welche für $0 < \overline{H} < 1$ mit der in Satz 8 vorkommenden algebraischen Gleichung äquivalent ist. Man kann zeigen, daß diese genau eine in $(0,1)$ liegende Wurzel besitzt, welche folglich mit H identisch ist. Ist durch $(a_1, a_2) = (u, v)$ ein optimaler Operator gekennzeichnet, so gilt für diesen wegen $(u,v) \in \Omega^*$ nach (14.17) wegen $\Psi = \overline{H}$

$$u = \varphi_0(\overline{H}), \quad v = X(u) = \chi(u, \overline{H}),$$

was in Satz 8 gerade behauptet wird. Da (u,ψ) ein innerer Punkt von $\mathfrak{M}$ ist, gilt $(u,v) \in \mathfrak{E}$. Der durch (u,v) bestimmte Operator aus $\mathfrak{F}_{k+1}$ ist also wegen Satz 2 auch absolut stabil. Die asymptotische Darstellung von $\bar{H}$ schließlich ist eine einfache Folge davon, daß $\bar{H}$ der im Satz genannten algebraischen Gleichung genügt.
Bisher war $\max_{\xi} H(a_1,a_2) < 1$ vorausgesetzt. Für $D_3 \geqq 52$ ist diese Voraussetzung überflüssig, weil dann - wie wir sahen - von selbst $\psi < 1$ ausfällt. Damit ist Satz 8 vollständig bewiesen.

$\log D_3$	$\bar{H}$ exakt	$\bar{H}$ asympt.
2.0	.85	1.2
2.5	.67	.87
3.0	.52	.65
3.5	.41	.49
4.0	.31	.37
4.5	.24	.27
5.0	.19	.21
5.5	.14	.15
6.0	.11	.12
6.5	.083	.087
7.0	.063	.065
7.5	.047	.049
8.0	.036	.037
8.5	.027	.027
9.0	.020	.020
9.5	.015	.015
10.0	.012	.012
10.5	.0086	.0087
11.0	.0065	.0065
11.5	.0049	.0049
12.0	.0037	.0037

Tafel II

Anmerkungen:
1) In Tafel II sind für verschiedene Werte von D_3 der exakte Wert von $\bar{H}$ und der aus der asymptotischen Darstellung gewonnene gegenübergestellt. Man sieht, daß die asymptotische Darstellung sehr gute Näherungen liefert.

2) Für $D_3 \to \infty$ gilt $\bar{H} \to 0$ und somit

$$u \to 1 , \quad v \to 0 .$$

Das bedeutet (vgl.auch(14.1)), daß die Simpson - Regel bezüglich der absolut stabilen Operatoren dritter Ordnung und vom Grade vier die gleiche Grenzstellung einnimmt wie im Falle $k = 2 , p = 3$. Man vergleiche hierzu Paragraph 13, Anmerkung 2. Wie man jedoch der Figur 5 entnimmt, hat $H(a_1,\chi(a_1)) = \psi(a_1)$ in $\mathfrak{M}_2$ eine sehr schwache Steigung. Ist $\bar{u}$ dadurch gekennzeichnet, daß $(\bar{u},\psi(\bar{u})) \in \mathfrak{E}_1$ gilt, und setzt man $\bar{v} = \chi(\bar{u})$, so steht der durch $(a_1,a_2) = (\bar{u},\bar{v})$ bestimmte Operator dem optimalen an Genauigkeit nicht viel nach. Für $D_3 \to \infty$ gilt $(\bar{u},\bar{v}) \to (0,0)$; man erhält also die durch $-s^3 + 1$ erzeugte Formel. Schließlich sind alle "zwischen" den so bestimmten Formeln, erzeugt durch das Polynom

$$-s^3 + \mu s + 1 - \mu \quad (0 \leqq \mu \leqq 1) ,$$

für $D_3 \to \infty$ als mit der Simpson - Regel gleichwertig zu erachten.

3) Satz 7 und Satz 8 legen einen Vergleich der optimalen Verfahren zweiter und dritter Ordnung nahe. Bei festem λ ist für $M_4 > 0$ stets $D_2 > D_3$. Für alle hinreichend kleinen $K > 0$ ist daher $\bar{H}$ im Falle $k = 2$ kleiner als im Falle $k = 3$.

§15. Optimale Operatoren für $h \to 0$.

In diesem Paragraphen werden nur qualitative Ergebnisse für $h \to 0$ erzielt. $h \to 0$ steht repräsentativ für "für alle hinreichend kleinen h". Es wird wieder $f_y(x,y) = g = \text{const}$ vorausgesetzt, jedoch wird nur $g \neq 0$ verlangt. Außerdem sei stets $k \geq 2$ und $L \in \mathcal{L}_{k+1}$. L werde zudem erzeugt durch ein Polynom

(15.1) $$\varrho(s) = \prod_{i=1}^{k} (s - s_i)$$

mit einfachen Wurzeln s_i $(i = 1,\dots,k)$, für welche

(15.2) $$s_1 = 1\,, \quad |s_i| \leq 1 \quad (i = 2,\dots,k)$$

gilt. Es soll nur über wenige Schrittweiten integriert werden, so daß man

$$-t_n = \tau = \text{const}$$

annehmen kann. Außerdem setzen wir $r_n \equiv 0$ voraus, um die Auswirkungen von Elementarstörungen in den Anfangswerten studieren zu können. Die Variationsgleichung (8.5) hat dann die Gestalt

(15.3) $$\sum_{\varkappa=0}^{k} (a_\varkappa + H b_\varkappa)\, \eta_{n+\varkappa} = \tau$$

und ihre Lösung:

(15.4) $$\eta_n = \sum_{i=1}^{k} A_i \sigma_i^n - A\,.$$

Hier sind die $\sigma_i = \sigma_i(H)$ als die Wurzeln der charakteristischen Gleichung

(15.5) $$\varrho(s) + H \cdot \sigma(s) = 0$$

definiert, für welche wegen (15.2)

(15.6) $$\sigma_i = s_i + \mathcal{O}(H) \quad (H \to 0) \quad \text{für } i = 1,\dots,k$$

gilt. Man prüft leicht nach, daß wegen (2.1) speziell

(15.6)a $$\sigma_1 = 1 + H + \mathcal{O}(H^2)$$

gilt.

Während A und die σ_i nur von H abhängen, gehen in die A_i auch die Anfangswert-Fehler $\eta_0, \ldots, \eta_{k-1}$ ein.
Durch Einsetzen von (15.4) in (15.3) findet man bei Berücksichtigung, daß die s_i (15.5) erfüllen und daß wegen $p > 1$ (2.1) gilt,

$$A = \frac{\tau}{H \cdot \varrho'(1)} \quad . \tag{15.7}$$

Ferner kann man durch geeignete Linearkomposition der Gleichungen

$$\eta_\varkappa = \sum_{i=1}^{k} A_i \sigma_i^\varkappa - A \qquad (\varkappa = 0, \ldots, k-1)$$

alle A_i für $i \neq \mu$ eliminieren; das läuft auf das Lösen einer Interpolationsaufgabe hinaus. Man findet

$$A_\mu = \frac{1}{\varrho'(\sigma_\mu)} \cdot \sum_{\varkappa=0}^{k-1} \Lambda_{\mu\varkappa}(H) \cdot (\eta_\varkappa + A) \qquad (\mu = 1, \ldots, k) \,, \tag{15.8}$$

wenn die $\Lambda_{\mu\varkappa}(H)$ definiert werden als die Koeffizienten der Polynome

$$\ell_\mu(s) = \frac{\Lambda \ldots k}{\Lambda \neq \mu} (s - \sigma_i) = \sum_{\varkappa=0}^{k-1} \Lambda_{\mu\varkappa}(H) \cdot s^\varkappa \,. \tag{15.9}$$

Aus (15.8), (15.7) und (15.9) erhält man für $H \to 0$

$$A_\mu \sim \frac{1}{\varrho'(s_\mu)} \cdot \sum_{\varkappa=0}^{k-1} \Lambda_{\mu\varkappa} \eta_\varkappa + \frac{\tau}{H} \cdot \frac{\ell_\mu(1)}{\varrho'(1)} \,,$$

wenn $\Lambda_{\mu\varkappa} = \Lambda_{\mu\varkappa}(0)$ gesetzt wird. Wegen

$$\ell_1(1) \sim \varrho'(1) \,,$$
$$\ell_\mu(1) \sim \frac{1 - \sigma_1}{1 - s_\mu} \cdot \varrho'(1) \sim \frac{H}{1 - s_\mu} \cdot \varrho'(1) \qquad (\mu = 2, \ldots, k)$$

- das letztere wegen (15.6)a - findet man schließlich

$$A_1 \sim \frac{1}{\varrho'(1)} \cdot \left\{ \sum_{\varkappa=0}^{k-1} \Lambda_{1\varkappa} \eta_\varkappa + \frac{\tau}{H} \right\}, \tag{15.10}$$

$$A_\mu \sim \frac{1}{\varrho'(s_\mu)} \cdot \left\{ \sum_{\varkappa=0}^{k-1} \Lambda_{\mu\varkappa} \eta_\varkappa + \frac{\tau}{1 - s_\mu} \right\} \qquad (\mu = 2, \ldots, k) \,, \tag{15.11}$$

jeweils für $H \to 0$.
Damit ist die Lösung (15.4) der Variationsgleichung (15.3) einer qualitativen Erörterung zugänglich gemacht worden, die nicht nur ihr Wachstum für $n \to \infty$ im Auge hat.

Es sei darauf hingewiesen, daß τ keineswegs mit dem Faktor $1/H$ in η_n eingeht, wie dies zunächst wegen (15.10) scheinen mag. Denn man erhält aus (15.7) und (15.10) bei festem n

$$A_1 \sigma_1^n - A \sim \frac{1}{\varrho'(1)} \left\{ \sum_{\varkappa=0}^{k-1} \Lambda_{1\varkappa} \eta_\varkappa \sigma_1^n + n \cdot \tau \right\}. \tag{15.12}$$

T.E.Hull und A.C.R.Newbery 1961 und 1962 haben auf der Suche nach Operatoren mit kleinem Integrationsfehler nur den Ausdruck (15.12)

erfaßt, die anderen Komponenten von η_n jedoch als vernachlässigbar angesehen. Sie kamen so zu der Ansicht, daß für einen Operator mit günstiger Fehlerfortpflanzung die Größe $|\varrho'(1)|$ maximal ausfallen müsse, mithin daß der durch $(1-s)(1+s)^{k-1}$ erzeugte Operator für $h \to 0$ den Vorzug verdiene. Dieser Operator gehört nicht zu $\overline{P}_{k+1}$; auch sind wir im Falle $k=3$ schon zu entgegengesetzter Ansicht gekommen (vgl. § 14 Anmerkung 2). Dem entnimmt man, daß die von Hull/Newbery vernachlässigten Komponenten von η_n keineswegs vernachlässigt werden dürfen.

E.Fehlberg 1961 nimmt den Ausdruck

$$|c_{k+2}/\varrho'(1)|$$

als Maß für die Güte eines Verfahrens.[1] Wegen $\tau \sim c_{k+2} \cdot h^{k+2}$ $(h \to 0)$ wird also im wesentlichen die Konstante A erfaßt (vgl. (15.7)). Es ist daher fraglich, ob das numerische Verhalten des Verfahrens auf diese Weise überhaupt charakterisiert wird.

Für die Verfahren von Hull/Newbery ist wohl $|\varrho'(1)|$ besonders groß, zugleich gilt jedoch $\varrho'(-1) = 0$. Bedenkt man, daß unter der Bedingung (15.2) stets

$$1 \leq M^* = \max_{\substack{1 \leq \mu \leq k \\ 0 \leq \varkappa \leq k-1}} |\Lambda_{\mu\varkappa}| < \binom{k-1}{[\frac{k-1}{2}]} \tag{15.13}$$

gilt, so erkennt man an (15.10) und (15.11), wie schädlich sich ein kleiner Wert $|\varrho'(s_\mu)|$ auf $|A_\mu|$ und damit auf η_n auswirken muß.

Soll für jeden Satz $(\eta_0, \ldots, \eta_{k-1})$ von Anfangswert-Fehlern mit $|\eta_\varkappa| \leq K$, K eine positive Konstante, $\varkappa = 0,1,\ldots,k-1$, $|\eta_k|$ nicht zu groß werden, so hat man dafür zu sorgen, daß keiner der Ausdrücke

$$\left|\frac{\Lambda_{\mu\varkappa}}{\varrho'(s_\mu)}\right| \qquad (\mu = 1,\ldots,k;\ \varkappa = 0,\ldots,k-1)$$

einen großen Wert annimmt. Will man also ein "Stabilitätsmaß" einführen, so bietet sich mit Rücksicht auf (15.13) als ein solches die Größe

$$M = \min_{1 \leq \mu \leq k} |\varrho'(s_\mu)|$$

an. Es gilt nun der Satz

---

1) So auch P.Henrici 1961, S.223.

Satz 9: Für D-stabile Operatoren L der Ordnung $k \geq 2$ gilt $M \leq k$, und es ist $M = k$ genau dann, wenn L durch $-s^k + 1$ erzeugt wird.

Beweis: Für D-stabile Operatoren L gilt

$$(s_1, \ldots, s_k) \in \mathfrak{K} = \{(s_1, \ldots, s_k) \mid |s_\mu| \leq 1 \ (\mu = 1, \ldots, k)\}.$$

Es genügt daher zu zeigen, daß $M \leq k$ in $\mathfrak{K}$ und daß $M = k$ nur für den Fall gilt, daß $\varrho(s) = -s^k + C$ mit $|C| = 1$.

Es sei $(s_1, \ldots, s_k) \in \mathfrak{K}$. Man setze $\varrho(s) = -\prod_{\mu=1}^{k}(s - s_\mu)$ und

$$P = P(s_1, \ldots, s_k) = \prod_{\mu=1}^{k} \varrho'(s_\mu).$$

Wiederholte Anwendung des Maximumprinzips bezüglich einer Variablen ergibt, daß $|P|$ sein Maximum bezüglich $\mathfrak{K}$ nur unter der Bedingung

$$s_\mu = e^{i\varphi_\mu} \qquad (\mu = 1, \ldots, k)$$

mit reellen φ_μ annimmt. Unter dieser Bedingung nimmt $P\bar{P}$ sein Maximum höchstens dann an, wenn

$$\frac{\partial}{\partial \varphi_\mu}(P\bar{P}) = 0 \quad (\mu = 1, \ldots, k) \text{ und } P \neq 0.$$

Diese Bedingung ist, wie man durch etwas Rechnung feststellt, gleichwertig mit den beiden Bedingungen

$$(k-1)\varrho'(s_\mu) - s_\mu \varrho''(s_\mu) = 0 \qquad (\mu = 1, \ldots, k),$$

$$s_\mu \neq s_\nu \quad \text{für } \mu \neq \nu,$$

aus denen sofort

$$(k-1)\varrho'(s) - s\varrho''(s) \equiv 0$$

folgt. Koeffizientenvergleich zeigt, daß $|P|^2$ und damit $|P|$ sein Maximum nur dann annimmt, wenn

$$\varrho(s) = -s^k + C \quad \text{mit } |C| = 1.$$

Für alle diese Polynome ist $|P| = k^k$, welches zugleich der Maximalwert von $|P|$ in $\mathfrak{K}$ ist; jedoch nur $-s^k + 1$ ist das erzeugende Polynom eines Operators L. Ist $p \geq 1$, so ist dieser D-stabil und es gilt $M = k$. Ist schließlich L ein D-stabiler Operator, der durch ein von $-s^k + 1$ verschiedenes Polynom erzeugt wird, so gilt

$$|P| = \prod_{\mu=1}^{k} |\varrho'(s_\mu)| < k^k,$$

woraus sofort $M < k$ folgt, was zu beweisen war.

Es ist von großer Wichtigkeit, daß für $\varrho(s) = -s^k + 1$ nicht nur M seinen Maximalwert, sondern zugleich M^* sein Minimum annimmt (vgl. (15.13)). Das bedeutet, daß die Koeffizienten in (15.10) und (15.11)

bei dieser Wahl des Operators besonders ausgewogen, das heißt betragsmäßig gleichmäßig klein sind. Im Falle $k \equiv 0(2)$ ist zugleich $c_{k+2} = 0$ (vgl.(5.7)), was einen betragsmäßig kleinen Formelfehler zufolge hat.
Wir sind somit zu der Überzeugung gelangt, daß für $h \to 0$ die durch $-s^k + 1$ erzeugten Operatoren $L \in \overline{\mathfrak{P}}_{k+1}$ besonders zu bevorzugen sind (Newton-Cotes-Formeln). Mit einigem Recht kann man sie "optimal für $h \to 0$" nennen. Während das oben erwähnte Verfahren von Hull/Newbery schon deshalb nicht unsere Zustimmung erhalten konnte, weil es nicht zu $\overline{\mathfrak{P}}_{k+1}$ gehört, ist der von uns erörterte Operator zwar auch nicht absolut stabil, jedoch absolut stabilen Operatoren benachbart. Zudem bestätigen die exakten Überlegungen, die wir in den Fällen $k = 2$ und $k = 3$ anstellen konnten, die mehr qualitativen Ergebnisse dieses Paragraphen. Man vergleiche hierzu § 13 Anmerkung 2 und § 14 Anmerkung 2.

Literaturverzeichnis.

ahlquist,G. Stability and error bounds in the numerical integration of ordinary differential equations, Math.Scand.4 (1956) 33-53 und Kungl.Tekniska Högskolans Handlingar 130 (1959).

ehlberg,E. Numerisch stabile Interpolationsformeln mit günstiger Fehlerfortpflanzung, Z.angew.Math.Mech. 41 (1961) 101-110.

amming,R.W. Stable predictor-corrector-methods for ordinary differential equations, J.Ass.Comp.Mach. 6 (1959) 37-47.

enrici,P. Theoretical and experimental studies on the accumulation of error, Proc.Int.Conf.on Inform.Proc. UNESCO Paris (1959) 36-44.

- Discrete variable methods in ordinary differential equations, New York 1962.

ull,T.F. u. Newbery,A.C.R. Error bounds for a family of three-point integration procedures, J.Soc.Ind.Appl.Math.7 (1959) 402-412.

- Integration procedures, which minimize propagated errors, J.Soc.Ind.Appl.Math.9 (1961) 31-47.

- Corrector formulas for multi-step integration methods, J.Soc.Ind.Appl.Math.10 (1962) 351-369.

amke,E. Differentialgleichungen, Leipzig 1962.

rückeberg,F. Zur numerischen Integration und Fehlererfassung bei Anfangswertaufgaben gewöhnlicher Differentialgleichungen, Diss. Bonn 1961.

Liniger,W. Zur Stabilität der numerischen Integrationsmethoden für Differentialgleichungen, Diss.Lausanne 1957.

Milne,W.E. u. Reynolds,R.R. Stability of a numerical solution of differential equations, J.Ass.Comp.Mach.6 (1959) 196-203 u. 7 (1960) 46-56.

Quade,W. Numerische Integration von gewöhnlichen Differentialgleichungen durch Interpolation nach Hermite, Z.angew.Math.Mech.37 (1957) 161-169.
Über die Stabilität numerischer Methoden zur Integration gewöhnlicher Differentialgleichungen 1.Ordnung, Z.angew.Math.Mech.39 (1959) 117-134.

Robertson,H.H. Some new formula for numerical integration of ordinary differential equations, Proc.Int.Conf.on Inform.Proc. UNESCO Paris 1959 106-108.

Rutishauser,H. Über die Instabilität von Methoden zur Integration gewöhnlicher Differentialgleichungen, Z.angew.Math.Phys. 3 (1952) 65-74.

Lebenslauf

Am 22. November 1933 wurde ich, Manfred Reimer, als Sohn des Oberbaurat Otto Reimer und seiner Ehefrau Erna in Breslau geboren. Ostern 1940 wurde ich dort in die Grundschule eingeschult. Nach der Vertreibung aus Schlesien fand ich 1945 in Hamburg eine zweite Heimat. 1953 bestand ich an der Albrecht-Thaer-Schule in Hamburg, einem naturwissenschaftlichen Gymnasium, die Reifeprüfung. Nach halbjährigem Baupraktikum begann ich im Winter 1953/54 mit dem Studium des Bauingenieurwesens an der Technischen Hochschule Hannover. Im Sommer 1954 änderte ich mein Studienziel und studierte die Fächer Mathematik und Physik an den Universitäten Göttingen und Tübingen. 1958 bestand ich in Tübingen das Staatsexamen. Anschließend besuchte ich das Studienseminar in Stuttgart, wo ich 1960 die Assessoren-Prüfung für das Höhere Lehramt ablegte. Seit August 1960 bin ich wissenschaftlicher Mitarbeiter am Rechenzentrum der Universität in Tübingen.

Foto - Druck:
Chemoprint oHG.
Gießen